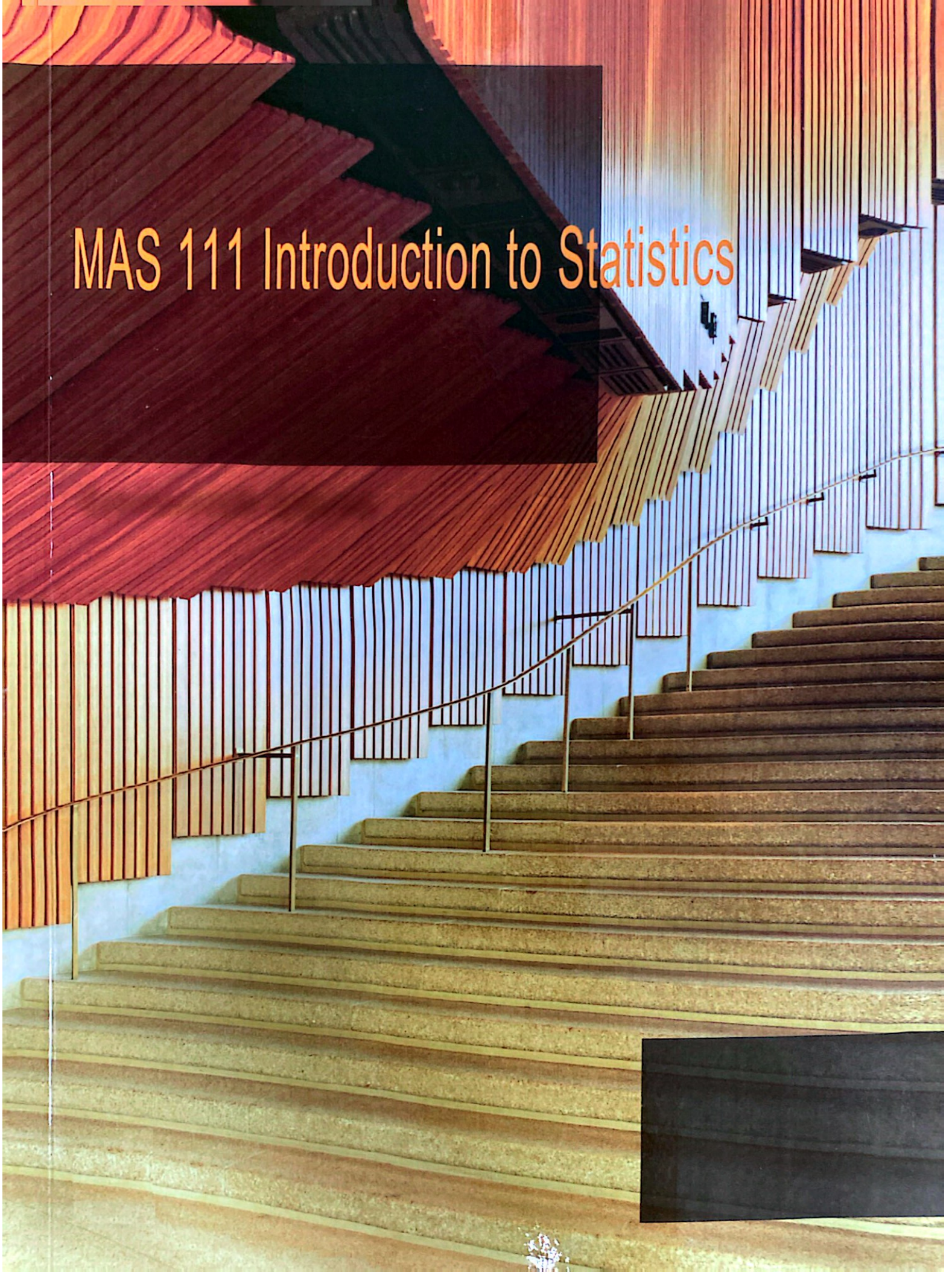


Scanned with CamScanner

Content

		Page
Chapter 1	What is Statistics?	1
Chapter 2	Describing Data: FREQUENCY TABLES, FREQUENCY DISTRIBUTIONS, AND GRAPHIC PRESENTATION	18
Chapter 3	Describing Data: NUMERICAL MEASURES	51
Chapter 4	Describing Data: DISPLAYING AND EXPLORING DATA	94
Chapter 5	A Survey of Probability Concepts	132
Chapter 13	Correlation and 13 Linear Regression	436

Reference

Statistical Techniques in Business & Economics

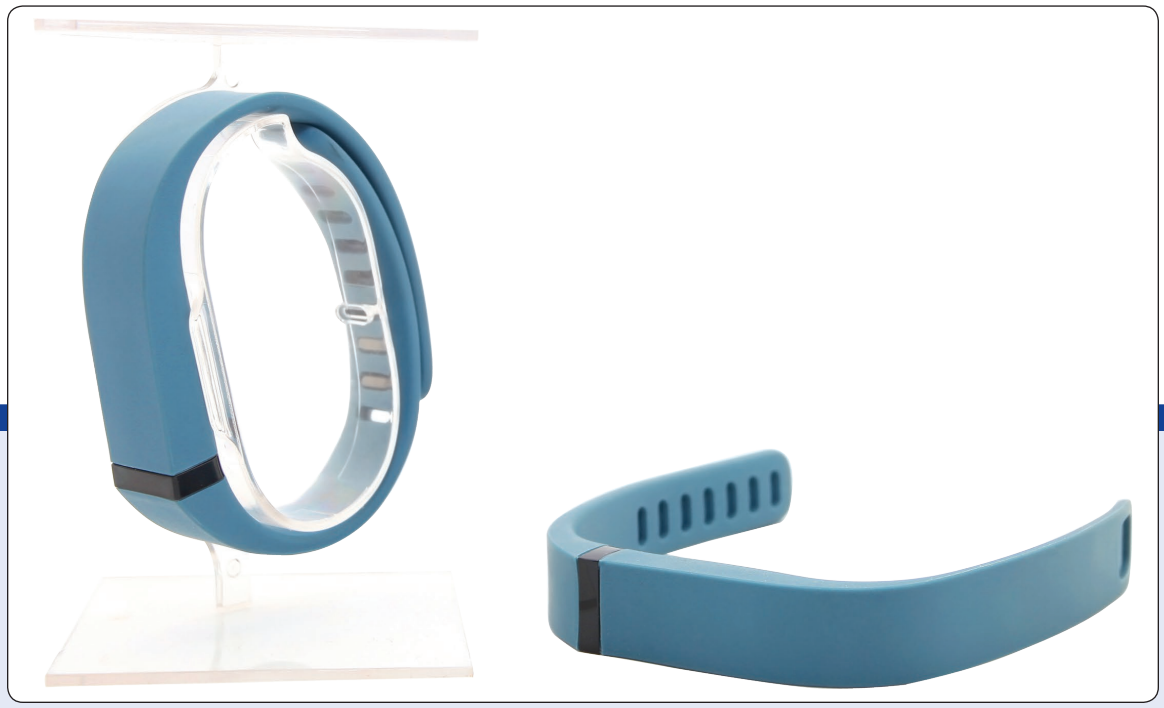
Seventeenth Edition

LIND MARCHAL WATHEN

Scanned with CamScanner

What is Statistics?

1



© Kelvin Wong/Shutterstock.com

▲ **BEST BUY** sells Fitbit wearable technology products that track a person's physical activity and sleep quality. The Fitbit technology collects daily information on a person's number of steps so that a person can track calories consumed. The information can be synced with a cell phone and displayed with a Fitbit app. Assume you know the daily number of Fitbit Flex 2 units sold last month at the Best Buy store in Collegeville, Pennsylvania. Describe a situation where the number of units sold is considered a sample. Illustrate a second situation where the number of units sold is considered a population. (See Exercise 11 and LO1-3.)

LEARNING OBJECTIVES

When you have completed this chapter, you will be able to:

LO1-1 Explain why knowledge of statistics is important.

LO1-2 Define statistics and provide an example of how statistics is applied.

LO1-3 Differentiate between descriptive and inferential statistics.

LO1-4 Classify variables as qualitative or quantitative, and discrete or continuous.

LO1-5 Distinguish between nominal, ordinal, interval, and ratio levels of measurement.

LO1-6 List the values associated with the practice of statistics.

© Gregor Schuster/Getty Images RF

INTRODUCTION

Suppose you work for a large company and your supervisor asks you to decide if a new version of a smartphone should be produced and sold. You start by thinking about the product's innovations and new features. Then, you stop and realize the consequences of the decision. The product will need to make a profit so the pricing and the costs of production and distribution are all very important. The decision to introduce the product is based on many alternatives. So how will you know? Where do you start?

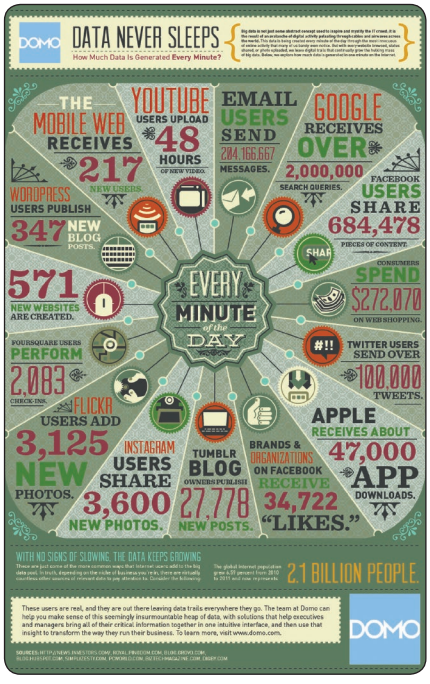
Without a long experience in the industry, beginning to develop an intelligence that will make you an expert is essential. You select three other people to work with and meet with them. The conversation focuses on what you need to know and what information and data you need. In your meeting, many questions are asked. How many competitors are already in the market? How are smartphones priced? What design features do competitors' products have? What features does the market require? What do customers want in a smartphone? What do customers like about the existing products? The answers will be based on business intelligence consisting of data and information collected through customer surveys, engineering analysis, and market research. In the end, your presentation to support your decision regarding the introduction of a new smartphone is based on the statistics that you use to summarize and organize your data, the statistics that you use to compare the new product to existing products, and the statistics to estimate future sales, costs, and revenues. The statistics will be the focus of the conversation that you will have with your supervisor about this very important decision.

As a decision maker, you will need to acquire and analyze data to support your decisions. The purpose of this text is to develop your knowledge of basic statistical techniques and methods and how to apply them to develop the business and personal intelligence that will help you make decisions.

LO1-1
Explain why knowledge of statistics is important.

WHY STUDY STATISTICS?

If you look through your university catalogue, you will find that statistics is required for many college programs. As you investigate a future career in accounting, economics, human resources, finance, business analytics, or other business area, you also will discover that statistics is required as part of these college programs. So why is statistics a requirement in so many disciplines?


DOMO
DATA NEVER SLEEPS
How Much Data Is Generated Every Minute?
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.
EVERY MINUTE of the DAY
WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING
2.1 BILLION PEOPLE
These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.
DOMO

A major driver of the requirement for statistics knowledge is the technologies available for capturing data. Examples include the technology that Google uses to track how Internet users access websites. As people use Google to search the Internet, Google records every search and then uses these data to sort and prioritize the results for future Internet searches. One recent estimate indicates that Google processes 20,000 terabytes of information per day. Big-box retailers like Target, Walmart, Kroger, and others scan every purchase and use the data to manage the distribution of products, to make decisions about marketing and sales, and to track daily and even hourly sales. Police departments collect and use data to provide city residents with maps that communicate information about crimes committed and their location. Every organization is collecting and using data to develop knowledge and intelligence that will help people make informed decisions, and to track the implementation of their decisions. The graphic to the left shows the amount of data generated every minute (www.domo.com). A good working knowledge of statistics is useful for summarizing and organizing data to provide information that is useful and supportive of decision making. Statistics is used to make valid comparisons and to predict the outcomes of decisions.

In summary, there are at least three reasons for studying statistics: (1) data are collected everywhere and require statistical knowledge to

make the information useful, (2) statistical techniques are used to make professional and personal decisions, and (3) no matter what your career, you will need a knowledge of statistics to understand the world and to be conversant in your career. An understanding of statistics and statistical method will help you make more effective personal and professional decisions.

LO1-2
Define statistics and provide an example of how statistics is applied.

WHAT IS MEANT BY STATISTICS?

This question can be rephrased in two, subtly different ways: what are statistics and what is statistics? To answer the first question, a statistic is a number used to communicate a piece of information. Examples of **statistics** are:

- The inflation rate is 2%.
- Your grade point average is 3.5.
- The price of a new Tesla Model S sedan is $79,570.

STATISTICS IN ACTION

A feature of our textbook is called *Statistics in Action*. Read each one carefully to get an appreciation of the wide application of statistics in management, economics, nursing, law enforcement, sports, and other disciplines.

- In 2015, *Forbes* published a list of the richest Americans. William Gates, founder of Microsoft Corporation, is the richest. His net worth is estimated at $76.0 billion. (www.forbes.com)
- In 2015, the four largest privately owned American companies, ranked by revenue, were Cargill, Koch Industries, Dell, and Albertsons. (www.forbes.com)
- In the United States, a typical high school graduate earns $668 per week, a typical college graduate with a bachelor's degree earns $1,101 per week, and a typical college graduate with a master's degree earns $1,326 per week. (www.bls.gov/emp/ep_chart_001.htm)

Each of these statistics is a numerical fact and communicates a very limited piece of information that is not very useful by itself. However, if we recognize that each of these statistics is part of a larger discussion, then the question "what **is** statistics" is applicable. Statistics is the set of knowledge and skills used to organize, summarize, and analyze data. The results of statistical analysis will start interesting conversations in the search for knowledge and intelligence that will help us make decisions. For example:

- The inflation rate for the calendar year was 0.7%. By applying statistics we could compare this year's inflation rate to the past observations of inflation. Is it higher, lower, or about the same? Is there a trend of increasing or decreasing inflation? Is there a relationship between interest rates and government bonds?
- Your grade point average (GPA) is 3.5. By collecting data and applying statistics, you can determine the required GPA to be admitted to the Master of Business Administration program at the University of Chicago, Harvard, or the University of Michigan. You can determine the likelihood that you would be admitted to a particular program. You may be interested in interviewing for a management position with Procter & Gamble. What GPA does Procter & Gamble require for college graduates with a bachelor's degree? Is there a range of acceptable GPAs?
- You are budgeting for a new car. You would like to own an electric car with a small carbon footprint. The price for the Tesla Model S Sedan is $79,570. By collecting additional data and applying statistics, you can analyze the alternatives. For example, another choice is a hybrid car that runs on both gas and electricity such as a 2015 Toyota Prius. It can be purchased for about $28,659. Another hybrid, the Chevrolet Volt, costs $33,995. What are the differences in the cars' specifications? What additional information can be collected and summarized so that you can make a good purchase decision?

Another example of using statistics to provide information to evaluate decisions is the distribution and market share of Frito-Lay products. Data are collected on each of the Frito-Lay product lines. These data include the market share and the pounds of product sold. Statistics is used to present this information in a bar chart in Chart 1–1. It clearly shows Frito-Lay's dominance in the potato, corn, and tortilla chip markets. It also shows the absolute measure of pounds of each product line consumed in the United States.

These examples show that statistics is more than the presentation of numerical information. Statistics is about collecting and processing information to create a conversation, to stimulate additional questions, and to provide a basis for making decisions. Specifically, we define **statistics** as:

STATISTICS The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

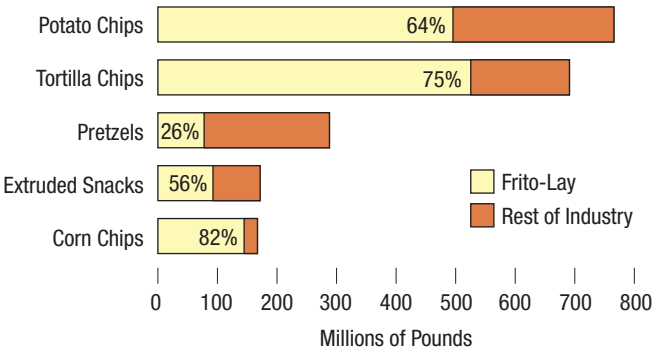


CHART 1–1 Frito-Lay Volume and Share of Major Snack Chip Categories in U.S. Supermarkets

In this book, you will learn the basic techniques and applications of statistics that you can use to support your decisions, both personal and professional. To start, we will differentiate between descriptive and inferential statistics.

LO1-3
Differentiate between descriptive and inferential statistics.

TYPES OF STATISTICS

When we use statistics to generate information for decision making from data, we use either descriptive statistics or inferential statistics. Their application depends on the questions asked and the type of data available.

Descriptive Statistics

Masses of unorganized data—such as the census of population, the weekly earnings of thousands of computer programmers, and the individual responses of 2,000 registered voters regarding their choice for president of the United States—are of little value as is. However, descriptive statistics can be used to organize data into a meaningful form. We define **descriptive statistics** as:

DESCRIPTIVE STATISTICS Methods of organizing, summarizing, and presenting data in an informative way.

The following are examples that apply descriptive statistics to summarize a large amount of data and provide information that is easy to understand.

- There are a total of 46,837 miles of interstate highways in the United States. The interstate system represents only 1% of the nation's total roads but carries more than 20% of the traffic. The longest is I-90, which stretches from Boston to Seattle, a distance of 3,099 miles. The shortest is I-878 in New York City, which is 0.70 mile in length. Alaska does not have any interstate highways, Texas has the most interstate miles at 3,232, and New York has the most interstate routes with 28.
- The average person spent $133.91 on traditional Valentine's Day merchandise in 2014. This is an increase of $2.94 from 2013. As in previous years, men spent more than twice the amount women spent on the holiday. The average man spent $108.38 to impress the people in his life while women only spent $48.41.

Statistical methods and techniques to generate descriptive statistics are presented in Chapters 2 and 4. These include organizing and summarizing data with frequency distributions and presenting frequency distributions with charts and graphs. In addition, statistical measures to summarize the characteristics of a distribution are discussed in Chapter 3.

Inferential Statistics

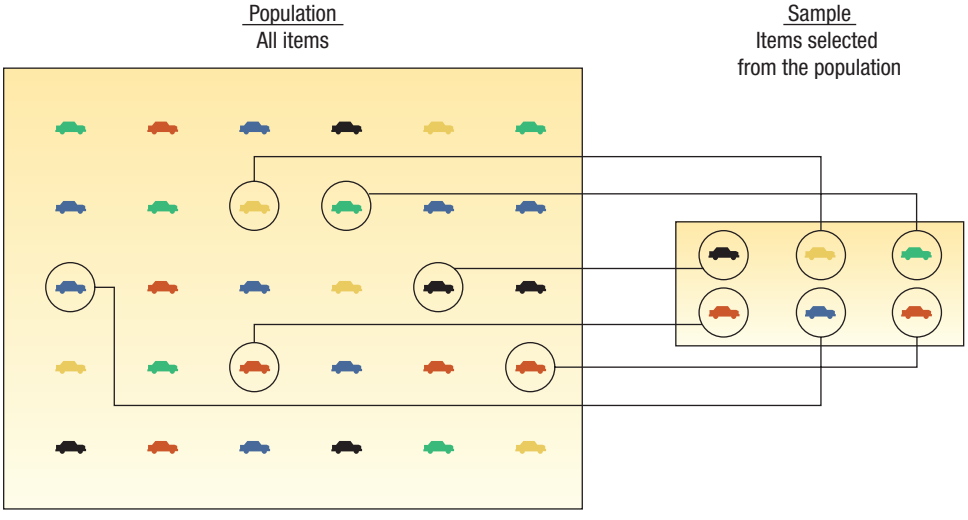
Sometimes we must make decisions based on a limited set of data. For example, we would like to know the operating characteristics, such as fuel efficiency measured by miles per gallon, of sport utility vehicles (SUVs) currently in use. If we spent a lot of time, money, and effort, all the owners of SUVs could be surveyed. In this case, our goal would be to survey the **population** of SUV owners.

POPULATION The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.

However, based on inferential statistics, we can survey a limited number of SUV owners and collect a **sample** from the population.

SAMPLE A portion, or part, of the population of interest.

Samples often are used to obtain reliable estimates of population parameters. (Sampling is discussed in Chapter 8.) In the process, we make trade-offs between the time, money, and effort to collect the data and the error of estimating a population parameter. The process of sampling SUVs is illustrated in the following graphic. In this example, we would like to know the mean or average SUV fuel efficiency. To estimate the mean of the population, six SUVs are sampled and the mean of their MPG is calculated.



So, the sample of six SUVs represents evidence from the population that we use to reach an inference or conclusion about the average MPG for all SUVs. The process of sampling from a population with the objective of estimating properties of a population is called **inferential statistics.**

INFERENTIAL STATISTICS The methods used to estimate a property of a population on the basis of a sample.

STATISTICS IN ACTION

Where did statistics get its start? In 1662 John Graunt published an article called "Natural and Political Observations Made upon Bills of Mortality." The author's "observations" were the result of a study and analysis of a weekly church publication called "Bill of Mortality," which listed births, christenings, and deaths and their causes. Graunt realized that the Bills of Mortality represented only a fraction of all births and deaths in London. However, he used the data to reach broad conclusions or inferences about the impact of disease, such as the plague, on the general population. His logic is an example of statistical inference. His analysis and interpretation of the data are thought to mark the start of statistics.

Inferential statistics is widely applied to learn something about a population in business, agriculture, politics, and government, as shown in the following examples:

- Television networks constantly monitor the popularity of their programs by hiring Nielsen and other organizations to sample the preferences of TV viewers. For example, 9.0% of a sample of households with TVs watched The Big Bang Theory during the week of November 2, 2015 (www.nielsen.com). These program ratings are used to make decisions about advertising rates and whether to continue or cancel a program.
- In 2015, a sample of U.S. Internal Revenue Service tax preparation volunteers were tested with three standard tax returns. The sample indicated that tax returns were completed with a 49% accuracy rate. In other words there were errors on about half of the returns. In this example, the statistics are used to make decisions about how to improve the accuracy rate by correcting the most common errors and improving the training of volunteers.

A feature of our text is self-review problems. There are a number of them interspersed throughout each chapter. The first self-review follows. Each self-review tests your comprehension of preceding material. The answer and method of solution are given in Appendix E. You can find the answer to the following self-review in 1–1 in Appendix E. We recommend that you solve each one and then check your answer.

SELF-REVIEW 1–1

The answers are in Appendix E.

The Atlanta-based advertising firm Brandon and Associates asked a sample of 1,960 consumers to try a newly developed chicken dinner by Boston Market. Of the 1,960 sampled, 1,176 said they would purchase the dinner if it is marketed.

(a) Is this an example of descriptive statistics or inferential statistics? Explain.
(b) What could Brandon and Associates report to Boston Market regarding acceptance of the chicken dinner in the population?

LO1-4
Classify variables as qualitative or quantitative, and discrete or continuous.

TYPES OF VARIABLES

There are two basic types of variables: (1) qualitative and (2) quantitative (see Chart 1–2). When an object or individual is observed and recorded as a nonnumeric characteristic, it is a qualitative variable or an attribute. Examples of qualitative variables are gender, beverage preference, type of vehicle owned, state of birth, and eye color. When a variable is qualitative, we usually count the number of observations for each category and determine

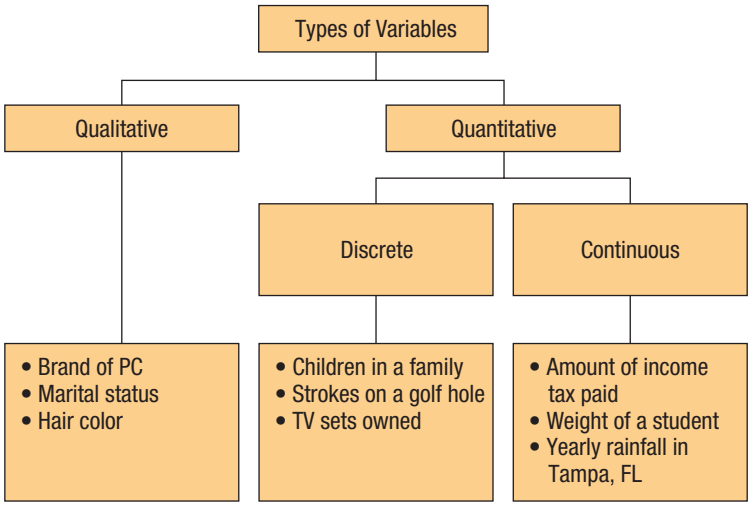


CHART 1–2 Summary of the Types of Variables

what percent are in each category. For example, if we observe the variable eye color, what percent of the population has blue eyes and what percent has brown eyes? If the variable is type of vehicle, what percent of the total number of cars sold last month were SUVs? Qualitative variables are often summarized in charts and bar graphs (Chapter 2).

When a variable can be reported numerically, it is called a quantitative variable. Examples of quantitative variables are the balance in your checking account, the number of gigabytes of data used on your cell phone plan last month, the life of a car battery (such as 42 months), and the number of people employed by a company.

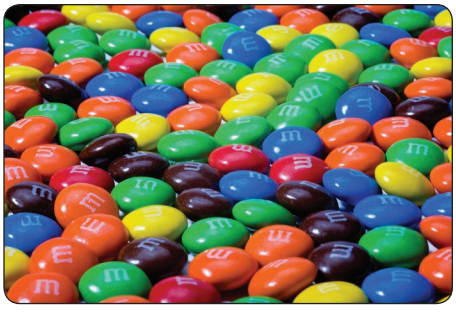
Quantitative variables are either discrete or continuous. Discrete variables can assume only certain values, and there are "gaps" between the values. Examples of discrete variables are the number of bedrooms in a house (1, 2, 3, 4, etc.), the number of cars arriving at Exit 25 on I-4 in Florida near Walt Disney World in an hour (326, 421, etc.), and the number of students in each section of a statistics course (25 in section A, 42 in section B, and 18 in section C). We count, for example, the number of cars arriving at Exit 25 on I-4, and we count the number of statistics students in each section. Notice that a home can have 3 or 4 bedrooms, but it cannot have 3.56 bedrooms. Thus, there is a "gap" between possible values. Typically, discrete variables are counted.

Observations of a continuous variable can assume any value within a specific range. Examples of continuous variables are the air pressure in a tire and the weight of a shipment of tomatoes. Other examples are the ounces of raisins in a box of raisin bran cereal and the duration of flights from Orlando to San Diego. Grade point average (GPA) is a continuous variable. We could report the GPA of a particular student as 3.2576952. The usual practice is to round to 3 places—3.258. Typically, continuous variables result from measuring.

LO1-5
Distinguish between nominal, ordinal, interval, and ratio levels of measurement.

LEVELS OF MEASUREMENT

Data can be classified according to levels of measurement. The level of measurement determines how data should be summarized and presented. It also will indicate the type of statistical analysis that can be performed. Here are two examples of the relationship between measurement and how we apply statistics. There are six colors of candies in a bag of M&Ms. Suppose we assign brown a value of 1, yellow 2, blue 3, orange 4, green 5, and red 6. What kind of variable is the color of an M&M? It is a qualitative variable. Suppose someone summarizes M&M color by adding the assigned color values, divides the sum by the number of M&Ms, and reports that the mean color is 3.56. How do we interpret this statistic? You are correct in concluding that it has no meaning as a measure of M&M color. As a qualitative variable, we can only report the count and percentage of each color in a bag of M&Ms. As a second example, in a high school track meet there are eight competitors in the 400-meter run. We report the order of finish and that the mean finish is 4.5. What does the mean finish tell us? Nothing! In both of these instances, we have not used the appropriate statistics for the level of measurement.



© Ron Buskirk/Alamy Stock Photo

There are four levels of measurement: nominal, ordinal, interval, and ratio. The lowest, or the most primitive, measurement is the nominal level. The highest is the ratio level of measurement.

Nominal-Level Data

For the **nominal level of measurement,** observations of a qualitative variable are measured and recorded as labels or names. The labels or names can only be classified and counted. There is no particular order to the labels.

NOMINAL LEVEL OF MEASUREMENT Data recorded at the nominal level of measurement is represented as labels or names. They have no order. They can only be classified and counted.

The classification of the six colors of M&M milk chocolate candies is an example of the nominal level of measurement. We simply classify the candies by color. There is no natural order. That is, we could report the brown candies first, the orange first, or any of the other colors first. Recording the variable gender is another example of the nominal level of measurement. Suppose we count the number of students entering a football game with a student ID and report how many are men and how many are women. We could report either the men or the women first. For the data measured at the nominal level, we are limited to counting the number in each category of the variable. Often, we convert these counts to percentages. For example, a random sample of M&M candies reports the following percentages for each color:

Color	Percent in a bag
Blue	24%
Green	20%
Orange	16%
Yellow	14%
Red	13%
Brown	13%

To process the data for a variable measured at the nominal level, we often numerically code the labels or names. For example, if we are interested in measuring the home state for students at East Carolina University, we would assign a student's home state of Alabama a code of 1, Alaska a code of 2, Arizona a 3, and so on. Using this procedure with an alphabetical listing of states, Wisconsin is coded 49 and Wyoming 50. Realize that the number assigned to each state is still a label or name. The reason we assign numerical codes is to facilitate counting the number of students from each state with statistical software. Note that assigning numbers to the states does not give us license to manipulate the codes as numerical information. Specifically, in this example, 1 + 2 = 3 corresponds to Alabama + Alaska = Arizona. Clearly, the nominal level of measurement does not permit any mathematical operation that has any valid interpretation.

Ordinal-Level Data

The next higher level of measurement is the **ordinal level.** For this level of measurement a qualitative variable or attribute is either ranked or rated on a relative scale.

ORDINAL LEVEL OF MEASUREMENT Data recorded at the ordinal level of measurement is based on a relative ranking or rating of items based on a defined attribute or qualitative variable. Variables based on this level of measurement are only ranked or counted.

Best Business Climate

1. Florida
2. Utah
3. Texas
4. Georgia
5. Indiana
6. Tennessee
7. Nebraska
8. North Carolina
9. Virginia
10. Washington

For example, many businesses make decisions about where to locate their facilities; in other words, where is the best place for their business? Business Facilities (**www.businessfacilities.com**) publishes a list of the top 10 states for the "best business climate." The 2016 rankings are shown to the left. They are based on the evaluation of many different factors, including the cost of labor, business tax climate, quality of life, transportation infrastructure, educated workforce, and economic growth potential.

This is an example of an ordinal scale because the states are ranked in order of best to worst business climate. That is, we know the relative order of the states based

on the attribute. For example, in 2016 Florida had the best business climate and Utah was second. Indiana was fifth, and that was better than Tennessee but not as good as Georgia. Notice we cannot say that Florida's business climate is five times better than Indiana's business climate because the magnitude of the differences between the states is not known. To put it another way, we do not know if the magnitude of the difference between Louisiana and Utah is the same as between Texas and Georgia.

Another example of the ordinal level measure is based on a scale that measures an attribute. This type of scale is used when students rate instructors on a variety of attributes. One attribute may be: "Overall, how do you rate the quality of instruction in this class?" A student's response is recorded on a relative scale of inferior, poor, good, excellent, and superior. An important characteristic of using a relative measurement scale is that we cannot distinguish the magnitude of the differences between groups. We do not know if the difference between "Superior" and "Good" is the same as the difference between "Poor" and "Inferior."

Table 1–1 lists the frequencies of 60 student ratings of instructional quality for Professor James Brunner in an Introduction to Finance course. The data are summarized based on the order of the scale used to rate the instructor. That is, they are summarized by the number of students who indicated a rating of superior (6), good (26), and so on. We also can convert the frequencies to percentages. About 43.3% (26/60) of the students rated the instructor as good.

TABLE 1–1 Rating of a Finance Professor

Rating	Frequency	Percentage
Superior	6	10.0%
Good	26	43.3%
Average	16	26.7%
Poor	9	15.0%
Inferior	3	5.0%

Interval-Level Data

The **interval level of measurement** is the next highest level. It includes all the characteristics of the ordinal level, but, in addition, the difference or interval between values is meaningful.

INTERVAL LEVEL OF MEASUREMENT For data recorded at the interval level of measurement, the interval or the distance between values is meaningful. The interval level of measurement is based on a scale with a known unit of measurement.

The Fahrenheit temperature scale is an example of the interval level of measurement. Suppose the high temperatures on three consecutive winter days in Boston are 28, 31, and 20 degrees Fahrenheit. These temperatures can be easily ranked, but we can also determine the interval or distance between temperatures. This is possible because 1 degree Fahrenheit represents a constant unit of measurement. That is, the distance between 10 and 15 degrees Fahrenheit is 5 degrees, and is the same as the 5-degree distance between 50 and 55 degrees Fahrenheit. It is also important to note that 0 is just a point on the scale. It does not represent the absence of the condition. The measurement of zero degrees Fahrenheit does not represent the absence of heat or cold. But by our own measurement scale, it is cold! A major limitation of a variable measured at the interval level is that we cannot make statements similar to 20 degrees Fahrenheit is twice as warm as 10 degrees Fahrenheit.

Another example of the interval scale of measurement is women's dress sizes. Listed below is information on several dimensions of a standard U.S. woman's dress.

Size	Bust (in)	Waist (in)	Hips (in)
8	32	24	35
10	34	26	37
12	36	28	39
14	38	30	41
16	40	32	43
18	42	34	45
20	44	36	47
22	46	38	49
24	48	40	51
26	50	42	53
28	52	44	55

Why is the "size" scale an interval measurement? Observe that as the size changes by two units (say from size 10 to size 12 or from size 24 to size 26), each of the measurements increases by 2 inches. To put it another way, the intervals are the same.

There is no natural zero point for dress size. A "size 0" dress does not have "zero" material. Instead, it would have a 24-inch bust, 16-inch waist, and 27-inch hips. Moreover, the ratios are not reasonable. If you divide a size 28 by a size 14, you do not get the same answer as dividing a size 20 by a size 10. Neither ratio is equal to two, as the "size" number would suggest. In short, if the distances between the numbers make sense, but the ratios do not, then you have an interval scale of measurement.

Ratio-Level Data

Almost all quantitative variables are recorded on the **ratio level of measurement.** The ratio level is the "highest" level of measurement. It has all the characteristics of the interval level, but, in addition, the 0 point and the ratio between two numbers are both meaningful.

RATIO LEVEL OF MEASUREMENT Data recorded at the ratio level of measurement are based on a scale with a known unit of measurement and a meaningful interpretation of zero on the scale.

Examples of the ratio scale of measurement include wages, units of production, weight, changes in stock prices, distance between branch offices, and height. Money is also a good illustration. If you have zero dollars, then you have no money, and a wage of $50 per hour is two times the wage of $25 per hour. Weight also is measured at the ratio level of measurement. If a scale is correctly calibrated, then it will read 0 when nothing is on the scale. Further, something that weighs 1 pound is half as heavy as something that weighs 2 pounds.

Table 1–2 illustrates the ratio scale of measurement for the variable, annual income for four father-and-son combinations. Observe that the senior Lahey earns twice as much as his son. In the Rho family, the son makes twice as much as the father.

TABLE 1–2 Father–Son Income Combinations

Name	Father	Son
Lahey	$80,000	$ 40,000
Nale	90,000	30,000
Rho	60,000	120,000
Steele	75,000	130,000

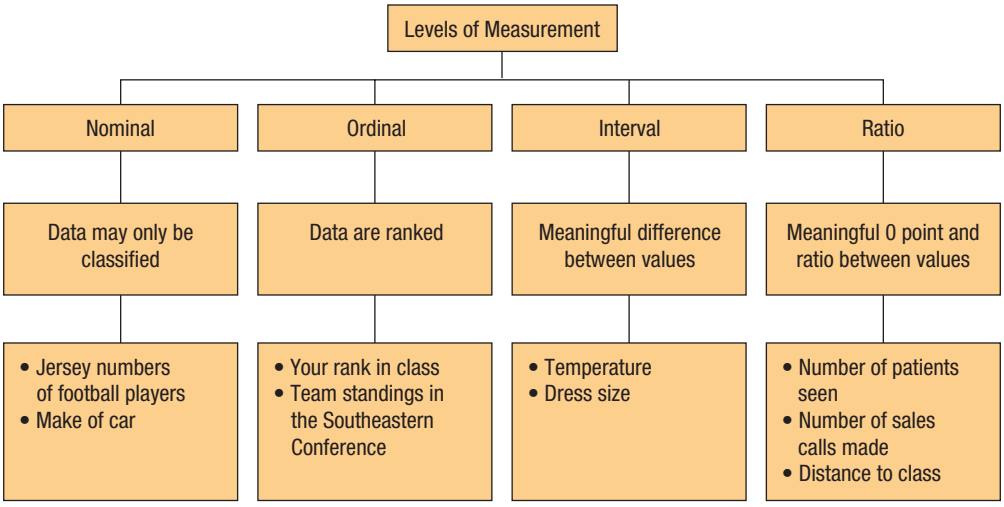


CHART 1–3 Summary and Examples of the Characteristics for Levels of Measurement

Chart 1–3 summarizes the major characteristics of the various levels of measurement. The level of measurement will determine the type of statistical methods that can be used to analyze a variable. Statistical methods to analyze variables measured on a nominal level are discussed in Chapter 15; methods for ordinal-level variables are discussed in Chapter 16. Statistical methods to analyze variables measured on an interval or ratio level are presented in Chapters 9 through 14.

SELF-REVIEW 1–2

(a) The mean age of people who listen to talk radio is 42.1 years. What level of measurement is used to assess the variable age?
(b) In a survey of luxury-car owners, 8% of the U.S. population owned luxury cars. In California and Georgia, 14% of people owned luxury cars. Two variables are included in this information. What are they and how are they measured?

EXERCISES

The answers to the odd-numbered exercises are in Appendix D.

1. What is the level of measurement for each of the following variables?
- **a.** Student IQ ratings.
- **b.** Distance students travel to class.
- **c.** The jersey numbers of a sorority soccer team.
- **d.** A student's state of birth.
- **e.** A student's academic class—that is, freshman, sophomore, junior, or senior.
- **f.** Number of hours students study per week.

2. *Slate* is a daily magazine on the Web. Its business activities can be described by a number of variables. What is the level of measurement for each of the following variables?
- **a.** The number of hits on their website on Saturday between 8:00 am and 9:00 am.
- **b.** The departments, such as food and drink, politics, foreign policy, sports, etc.
- **c.** The number of weekly hits on the Sam's Club ad.
- **d.** The number of years each employee has been employed with Slate.

3. On the Web, go to your favorite news source and find examples of each type of variable. Write a brief memo that lists the variables and describes them in terms of qualitative or quantitative, discrete or continuous, and the measurement level.

4. For each of the following, determine whether the group is a sample or a population.
- **a.** The participants in a study of a new cholesterol drug.
- **b.** The drivers who received a speeding ticket in Kansas City last month.
- **c.** People on welfare in Cook County (Chicago), Illinois.
- **d.** The 30 stocks that make up the Dow Jones Industrial Average.

LO1-6
List the values associated with the practice of statistics.

ETHICS AND STATISTICS

Following events such as Wall Street money manager Bernie Madoff's Ponzi scheme, which swindled billions from investors, and financial misrepresentations by Enron and Tyco, business students need to understand that these events were based on the misrepresentation of business and financial information. In each case, people within each organization reported financial information to investors that indicated the companies were performing much better than the actual situation. When the true financial information was reported, the companies were worth much less than advertised. The result was many investors lost all or nearly all of the money they had invested.

The article "Statistics and Ethics: Some Advice for Young Statisticians," in *The American Statistician* 57, no. 1 (2003), offers guidance. The authors advise us to practice statistics with integrity and honesty, and urge us to "do the right thing" when collecting, organizing, summarizing, analyzing, and interpreting numerical information. The real contribution of statistics to society is a moral one. Financial analysts need to provide information that truly reflects a company's performance so as not to mislead individual investors. Information regarding product defects that may be harmful to people must be analyzed and reported with integrity and honesty. The authors of *The American Statistician* article further indicate that when we practice statistics, we need to maintain "an independent and principled point-of-view" when analyzing and reporting findings and results.

As you progress through this text, we will highlight ethical issues in the collection, analysis, presentation, and interpretation of statistical information. We also hope that, as you learn about using statistics, you will become a more informed consumer of information. For example, you will question a report based on data that do not fairly represent the population, a report that does not include all relevant statistics, one that includes an incorrect choice of statistical measures, or a presentation that introduces bias in a deliberate attempt to mislead or misrepresent.

BASIC BUSINESS ANALYTICS

A knowledge of statistics is necessary to support the increasing need for companies and organizations to apply business analytics. Business analytics is used to process and analyze data and information to support a story or narrative of a company's business, such as "what makes us profitable," "how will our customers respond to a change in marketing"? In addition to statistics, an ability to use computer software to summarize, organize, analyze, and present the findings of statistical analysis is essential. In this text, we will be using very elementary applications of business analytics using common and available computer software. Throughout our text, we will use Microsoft Excel and, occasionally, Minitab. Universities and colleges usually offer access to Microsoft Excel. Your computer already may be packaged with Microsoft Excel. If not, the Microsoft Office package with Excel often is sold at a reduced academic price through your university or college. In this text, we use Excel for the majority of the applications. We also use an Excel "Add-in" called MegaStat. If your instructor requires this package, it is available at **www.mhhe.com/megastat**. This add-in gives Excel the capability to produce additional statistical reports. Occasionally, we use Minitab to illustrate an application. See **www.minitab.com** for further information. Minitab also offers discounted academic pricing. The 2016 version of Microsoft Excel supports the analyses in our text. However,

earlier versions of Excel for Apple Mac computers do not have the necessary add-in. If you do not have Excel 2016 and are using an Apple Mac computer with Excel, you can download the free, trial version of Stat Plus at **www.analystsoft.com**. It is a statistical software package that will integrate with Excel for Mac computers.

The following example shows the application of Excel to perform a statistical summary. It refers to sales information from the Applewood Auto Group, a multi-location car sales and service company. The Applewood information has sales information for 180 vehicle sales. Each sale is described by several variables: the age of the buyer, whether the buyer is a repeat customer, the location of the dealership for the sale, the type of vehicle sold, and the profit for the sale. The following shows Excel's summary of statistics for the variable profit. The summary of profit shows the mean profit per vehicle was $1,843.17, the median profit was slightly more at $1,882.50, and profit ranged from $294 to $3,292.

APPLEWOOD AUTO GROUP.xlsx

	A	B	C	D	E	F	G	H
1	Age	Profit	Location	Vehicle-Type	Previous		*Profit*	
2	33	$1,889	Olean	SUV	1			
3	47	$1,461	Kane	Sedan	0		Mean	1843.17
4	44	$1,532	Tionesta	SUV	3		Standard Error	47.97
5	53	$1,220	Olean	Sedan	0		Median	1882.50
6	51	$1,674	Sheffield	Sedan	1		Mode	1915.00
7	41	$2,389	Kane	Truck	1		Standard Deviation	643.63
8	58	$2,058	Kane	SUV	1		Sample Variance	414256.61
9	35	$1,919	Tionesta	SUV	1		Kurtosis	-0.22
10	45	$1,266	Olean	Sedan	0		Skewness	-0.24
11	54	$2,991	Tionesta	Sedan	0		Range	2998
12	56	$2,695	Kane	Sedan	2		Minimum	294
13	41	$2,165	Tionesta	SUV	0		Maximum	3292
14	38	$1,766	Sheffield	SUV	0		Sum	331770
15	48	$1,952	Tionesta	Compact	1		Count	180

Throughout the text, we will motivate the use of computer software to summarize, describe, and present information and data. The applications of Excel are supported by instructions so that you can learn how to apply Excel to do statistical analysis. The instructions are presented in Appendix C of this text. These data and other data sets and files are available on the text's student website, **www.mhhe.com/lind17e**.

CHAPTER SUMMARY

I. Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

II. There are two types of statistics.
- **A.** Descriptive statistics are procedures used to organize and summarize data.
- **B.** Inferential statistics involve taking a sample from a population and making estimates about a population based on the sample results.
  - **1.** A population is an entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.
  - **2.** A sample is a part of the population.

III. There are two types of variables.
- **A.** A qualitative variable is nonnumeric.
  - **1.** Usually we are interested in the number or percent of the observations in each category.
  - **2.** Qualitative data are usually summarized in graphs and bar charts.

B. There are two types of quantitative variables and they are usually reported numerically.
  1. Discrete variables can assume only certain values, and there are usually gaps between values.
  2. A continuous variable can assume any value within a specified range.

IV. There are four levels of measurement.
A. With the nominal level, the data are sorted into categories with no particular order to the categories.
B. The ordinal level of measurement presumes that one classification is ranked higher than another.
C. The interval level of measurement has the ranking characteristic of the ordinal level of measurement plus the characteristic that the distance between values is a constant size.
D. The ratio level of measurement has all the characteristics of the interval level, plus there is a 0 point and the ratio of two values is meaningful.

CHAPTER EXERCISES

5. Explain the difference between qualitative and quantitative variables. Give an example of qualitative and quantitative variables.
6. Explain the difference between a sample and a population.
7. Explain the difference between a discrete and a continuous variable. Give an example of each not included in the text.
8. For the following situations, would you collect information using a sample or a population? Why?
  a. Statistics 201 is a course taught at a university. Professor Rauch has taught nearly 1,500 students in the course over the past 5 years. You would like to know the average grade for the course.
  b. As part of a research project, you need to report the average profit as a percentage of revenue for the #1-ranked corporation in the Fortune 500 for each of the last 10 years.
  c. You are looking forward to graduation and your first job as a salesperson for one of five large pharmaceutical corporations. Planning for your interviews, you will need to know about each company's mission, profitability, products, and markets.
  d. You are shopping for a new MP3 music player such as the Apple iPod. The manufacturers advertise the number of music tracks that can be stored in the memory. Usually, the advertisers assume relatively short, popular songs to estimate the number of tracks that can be stored. You, however, like Broadway musical tunes and they are much longer. You would like to estimate how many Broadway tunes will fit on your MP3 player.
9. Exits along interstate highways were formerly numbered successively from the western or southern border of a state. However, the Department of Transportation has recently changed most of them to agree with the numbers on the mile markers along the highway.
  a. What level of measurement were data on the consecutive exit numbers?
  b. What level of measurement are data on the milepost numbers?
  c. Discuss the advantages of the newer system.
10. A poll solicits a large number of college undergraduates for information on the following variables: the name of their cell phone provider (AT&T, Verizon, and so on), the numbers of minutes used last month (200, 400, for example), and their satisfaction with the service (Terrible, Adequate, Excellent, and so forth). What is the level of measurement for each of these three variables?
11. Best Buy sells Fitbit wearable technology products that track a person's activity. For example, the Fitbit technology collects daily information on a person's number of steps so that a person can track calories consumed. The information can be synced with a cell phone and displayed with a Fitbit app. Assume you know the daily number of Fitbit Flex

2 units sold last month at the Best Buy store in Collegeville, Pennsylvania. Describe a situation where the number of units sold is considered a sample. Illustrate a second situation where the number of units sold is considered a population.

12. Using the concepts of sample and population, describe how a presidential election is unlike an "exit" poll of the electorate.

13. Place these variables in the following classification tables. For each table, summarize your observations and evaluate if the results are generally true. For example, salary is reported as a continuous quantitative variable. It is also a continuous ratio-scaled variable.

a. Salary
b. Gender
c. Sales volume of MP3 players
d. Soft drink preference
e. Temperature
f. SAT scores
g. Student rank in class
h. Rating of a finance professor
i. Number of home video screens

	Discrete Variable	Continuous Variable
Qualitative		
Quantitative		a. Salary

	Discrete	Continuous
Nominal		
Ordinal		
Interval		
Ratio		a. Salary

14. Using data from such publications as the *Statistical Abstract of the United States, Forbes,* or any news source, give examples of variables measured with nominal, ordinal, interval, and ratio scales.

15. The Struthers Wells Corporation employs more than 10,000 white-collar workers in its sales offices and manufacturing facilities in the United States, Europe, and Asia. A sample of 300 U.S. workers revealed 120 would accept a transfer to a location outside the United States. On the basis of these findings, write a brief memo to Ms. Wanda Carter, Vice President of Human Services, regarding all white-collar workers in the firm and their willingness to relocate.

16. AVX Home Entertainment, Inc., recently began a "no-hassles" return policy. A sample of 500 customers who recently returned items showed 400 thought the policy was fair, 32 thought it took too long to complete the transaction, and the rest had no opinion. On the basis of this information, make an inference about customer reaction to the new policy.

17. FILE *The Wall Street Journal*'s website, **www.wsj.com**, reported the number of cars and light-duty trucks sold through October of 2014 and October of 2015. The top sixteen manufacturers are listed here. Sales data often is reported in this way to compare current sales to last year's sales.

Manufacturer	Year-to-Date Sales Through October 2015	Year-to-Date Sales Through October 2014
General Motors Corp.	2,562,840	2,434,707
Ford Motor Company	2,178,587	2,065,612
Toyota Motor Sales USA Inc.	2,071,446	1,975,368
Chrysler	1,814,268	1,687,313
American Honda Motor Co Inc.	1,320,217	1,281,777
Nissan North America Inc.	1,238,535	1,166,389
Hyundai Motor America	638,195	607,539
Kia Motors America Inc.	526,024	489,711
Subaru of America Inc.	480,331	418,497
Volkswagen of America Inc.	294,602	301,187
Mercedes-Benz	301,915	281,728
BMW of North America Inc.	279,395	267,193
Mazda Motor of America Inc.	267,158	259,751
Audi of America Inc.	165,103	146,133
Mitsubishi Motors N A, Inc.	80,683	64,564
Volvo	53,803	47,823

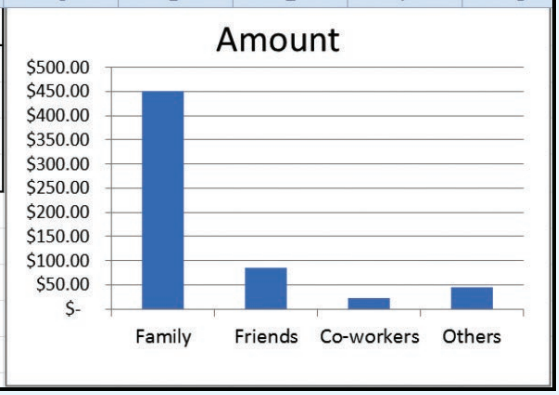
a. Using computer software, compare the October 2015 sales to the October 2014 sales for each manufacturer by computing the difference. Make a list of the manufacturers that increased sales compared to 2014; make a list of manufacturers that decreased sales.

b. Using computer software, compare 2014 sales to 2015 sales for each manufacturer by computing the percentage change in sales. Make a list of the manufacturers in order of increasing percentage changes. Which manufacturers are in the top five in percentage change? Which manufacturers are in the bottom five in percentage change?

c. Using computer software, first sort the data using the 2015 year-to-date sales. Then, design a bar graph to illustrate the 2014 and 2015 year-to-date sales for the top 12 manufacturers. Also, design a bar graph to illustrate the percentage change for the top 12 manufacturers. Compare these two graphs and prepare brief written comments.

18. The following chart depicts the average amounts spent by consumers on holiday gifts.

	A	B
1	Holiday Spending	Amount
2	Family	$451.34
3	Friends	$ 85.60
4	Co-workers	$ 22.40
5	Others	$ 44.52
6		
7		
8		
9		
10		

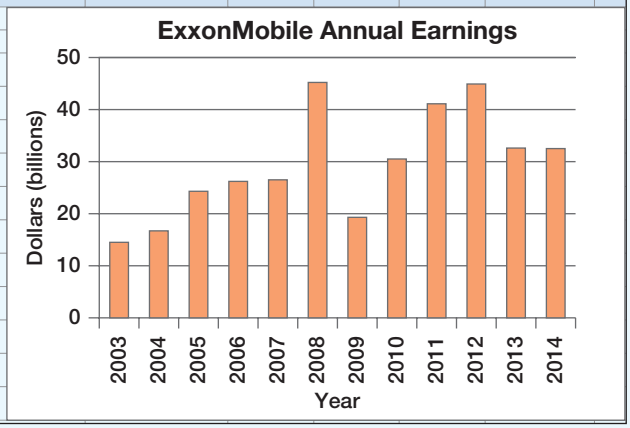


Write a brief report summarizing the amounts spent during the holidays. Be sure to include the total amount spent and the percent spent by each group.

19. The following chart depicts the earnings in billions of dollars for ExxonMobil for the period 2003 until 2014. Write a brief report discussing the earnings at ExxonMobil during

the period. Was one year higher than the others? Did the earnings increase, decrease, or stay the same over the period?

	A	B
1	Year	Earnings ($ billions)
2	2003	14.5
3	2004	16.7
4	2005	24.3
5	2006	26.2
6	2007	26.5
7	2008	45.2
8	2009	19.3
9	2010	30.5
10	2011	41.1
11	2012	44.9
12	2013	32.6
13	2014	32.5



DATA ANALYTICS

20. FILE Refer to the North Valley Real Estate data, which report information on homes sold in the area last year. Consider the following variables: selling price, number of bedrooms, township, and mortgage type.
- **a.** Which of the variables are qualitative and which are quantitative?
- **b.** How is each variable measured? Determine the level of measurement for each of the variables.

21. FILE Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Consider the following variables: number of wins, payroll, season attendance, whether the team is in the American or National League, and the number of home runs hit.
- **a.** Which of these variables are quantitative and which are qualitative?
- **b.** Determine the level of measurement for each of the variables.

22. FILE Refer to the Lincolnville School District bus data, which report information on the school district's bus fleet.
- **a.** Which of the variables are qualitative and which are quantitative?
- **b.** Determine the level of measurement for each variable.

2 Describing Data:

FREQUENCY TABLES, FREQUENCY DISTRIBUTIONS, AND GRAPHIC PRESENTATION

© rido/123RF

▲ **MERRILL LYNCH** recently completed a study of online investment portfolios for a sample of clients. For the 70 participants in the study, organize these data into a frequency distribution. (See Exercise 43 and LO2-3.)

LEARNING OBJECTIVES

When you have completed this chapter, you will be able to:

LO2-1 Summarize qualitative variables with frequency and relative frequency tables.

LO2-2 Display a frequency table using a bar or pie chart.

LO2-3 Summarize quantitative variables with frequency and relative frequency distributions.

LO2-4 Display a frequency distribution using a histogram or frequency polygon.

INTRODUCTION

The United States automobile retailing industry is highly competitive. It is dominated by megadealerships that own and operate 50 or more franchises, employ over 10,000 people, and generate several billion dollars in annual sales. Many of the top dealerships are publicly owned with shares traded on the New York Stock Exchange or NASDAQ. In 2014, the largest megadealership was AutoNation (ticker symbol AN), followed by Penske Auto Group (PAG), Group 1 Automotive, Inc. (ticker symbol GPI), and the privately owned Van Tuyl Group.

© Justin Sullivan/Getty Images

These large corporations use statistics and analytics to summarize and analyze data and information to support their decisions. As an example, we will look at the Applewood Auto group. It owns four dealerships and sells a wide range of vehicles. These include the popular Korean brands Kia and Hyundai, BMW and Volvo sedans and luxury SUVs, and a full line of Ford and Chevrolet cars and trucks.

Ms. Kathryn Ball is a member of the senior management team at Applewood Auto Group, which has its corporate offices adjacent to Kane Motors. She is responsible for tracking and analyzing vehicle sales and the profitability of those vehicles. Kathryn would like to summarize the profit earned on the vehicles sold with tables, charts, and graphs that she would review monthly. She wants to know the profit per vehicle sold, as well as the lowest and highest amount of profit. She is also interested in describing the demographics of the buyers. What are their ages? How many vehicles have they previously purchased from one of the Applewood dealerships? What type of vehicle did they purchase?

The Applewood Auto Group operates four dealerships:

- ***Tionesta Ford Lincoln*** sells Ford and Lincoln cars and trucks.
- ***Olean Automotive Inc.*** has the Nissan franchise as well as the General Motors brands of Chevrolet, Cadillac, and GMC Trucks.
- ***Sheffield Motors Inc.*** sells Buick, GMC trucks, Hyundai, and Kia.
- ***Kane Motors*** offers the Chrysler, Dodge, and Jeep line as well as BMW and Volvo.

APPLEWOOD AUTO GROUP

	A	B	C	D	E
1	Age	Profit	Location	Vehicle-Type	Previous
2	21	$1,387	Tionesta	Sedan	0
3	23	$1,754	Sheffield	SUV	1
4	24	$1,817	Sheffield	Hybrid	1
5	25	$1,040	Sheffield	Compact	0
6	26	$1,273	Kane	Sedan	1
7	27	$1,529	Sheffield	Sedan	1
8	27	$3,082	Kane	Truck	0
9	28	$1,951	Kane	SUV	1
10	28	$2,692	Tionesta	Compact	0
11	29	$1,206	Sheffield	Sedan	0
12	29	$1,342	Kane	Sedan	2
13	30	$443	Kane	Sedan	3
14	30	$754	Olean	Sedan	2
15	30	$1,621	Sheffield	Truck	1

Every month, Ms. Ball collects data from each of the four dealerships and enters them into an Excel spreadsheet. Last month the Applewood Auto Group sold 180 vehicles at the four dealerships. A copy of the first few observations appears to the left. The variables collected include:

- ***Age***—the age of the buyer at the time of the purchase.
- ***Profit***—the amount earned by the dealership on the sale of each vehicle.
- ***Location***—the dealership where the vehicle was purchased.
- ***Vehicle type***—SUV, sedan, compact, hybrid, or truck.
- ***Previous***—the number of vehicles previously purchased at any of the four Applewood dealerships by the consumer.

The entire data set is available at the McGraw-Hill website (**www.mhhe.com/lind17e**) and in Appendix A.4 at the end of the text.

LO2-1
Summarize qualitative variables with frequency and relative frequency tables.

CONSTRUCTING FREQUENCY TABLES

Recall from Chapter 1 that techniques used to describe a set of data are called descriptive statistics. Descriptive statistics organize data to show the general pattern of the data, to identify where values tend to concentrate, and to expose extreme or unusual data values. The first technique we discuss is a **frequency table.**

FREQUENCY TABLE A grouping of qualitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

In Chapter 1, we distinguished between qualitative and quantitative variables. To review, a qualitative variable is nonnumeric, that is, it can only be classified into distinct categories. Examples of qualitative data include political affiliation (Republican, Democrat, Independent, or other), state of birth (Alabama, . . . , Wyoming), and method of payment for a purchase at Barnes & Noble (cash, digital wallet, debit, or credit). On the other hand, quantitative variables are numerical in nature. Examples of quantitative data relating to college students include the price of their textbooks, their age, and the number of credit hours they are registered for this semester.

In the Applewood Auto Group data set, there are five variables for each vehicle sale: age of the buyer, amount of profit, dealer that made the sale, type of vehicle sold, and number of previous purchases by the buyer. The dealer and the type of vehicle are *qualitative* variables. The amount of profit, the age of the buyer, and the number of previous purchases are *quantitative* variables.

© Steve Cole/Getty Images RF

Suppose Ms. Ball wants to summarize last month's sales by location. The first step is to sort the vehicles sold last month according to their location and then tally, or count, the number sold at each location of the four locations: Tionesta, Olean, Sheffield, or Kane. The four locations are used to develop a frequency table with four mutually exclusive (distinctive) classes. Mutually exclusive classes means that a particular vehicle can be assigned to only one class. In addition, the frequency table must be collectively exhaustive. That is every vehicle sold last month is accounted for in the table. If every vehicle is included in the frequency table, the table will be collectively exhaustive and the total number of vehicles will be 180. How do we obtain these counts? Excel provides a tool called a Pivot Table that will quickly and accurately establish the four classes and do the counting. The Excel results follow in Table 2–1. The table shows a total of 180 vehicles and, of the 180 vehicles, 52 were sold at Kane Motors.

TABLE 2–1 Frequency Table for Vehicles Sold Last Month at Applewood Auto Group by Location

Location	Number of Cars
Kane	52
Olean	40
Sheffield	45
Tionesta	43
Total	180

Relative Class Frequencies

You can convert class frequencies to relative class frequencies to show the fraction of the total number of observations in each class. A relative frequency captures the relationship between a class frequency and the total number of observations. In the vehicle sales example, we may want to know the percentage of total cars sold at each of the four locations. To convert a frequency table to a relative frequency table, each of the class frequencies is divided by the total number of observations. Again, this is easily accomplished using Excel. The fraction of vehicles sold last month at the Kane location is 0.289, found by 52 divided by 180. The relative frequency for each location is shown in Table 2–2.

TABLE 2–2 Relative Frequency Table of Vehicles Sold by Location Last Month at Applewood Auto Group

Location	Number of Cars	Relative Frequency	Found by
Kane	52	.289	52/180
Olean	40	.222	40/180
Sheffield	45	.250	45/180
Tionesta	43	.239	43/180
Total	180	1.000	

LO2-2
Display a frequency table using a bar or pie chart.

GRAPHIC PRESENTATION OF QUALITATIVE DATA

The most common graphic form to present a qualitative variable is a **bar chart.** In most cases, the horizontal axis shows the variable of interest. The vertical axis shows the frequency or fraction of each of the possible outcomes. A distinguishing feature of a bar chart is there is distance or a gap between the bars. That is, because the variable of interest is qualitative, the bars are not adjacent to each other. Thus, a bar chart graphically describes a frequency table using a series of uniformly wide rectangles, where the height of each rectangle is the class frequency.

BAR CHART A graph that shows qualitative classes on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are proportional to the heights of the bars.

We use the Applewood Auto Group data as an example (Chart 2–1). The variables of interest are the location where the vehicle was sold and the number of vehicles sold at each location. We label the horizontal axis with the four locations and scale the vertical axis with the number sold. The variable location is of nominal scale, so the order of the locations on the horizontal axis does not matter. In Chart 2–1, the locations are listed alphabetically. The locations could also be in order of decreasing or increasing frequencies.

The height of the bars, or rectangles, corresponds to the number of vehicles at each location. There were 52 vehicles sold last month at the Kane location, so the height of the Kane bar is 52; the height of the bar for the Olean location is 40.

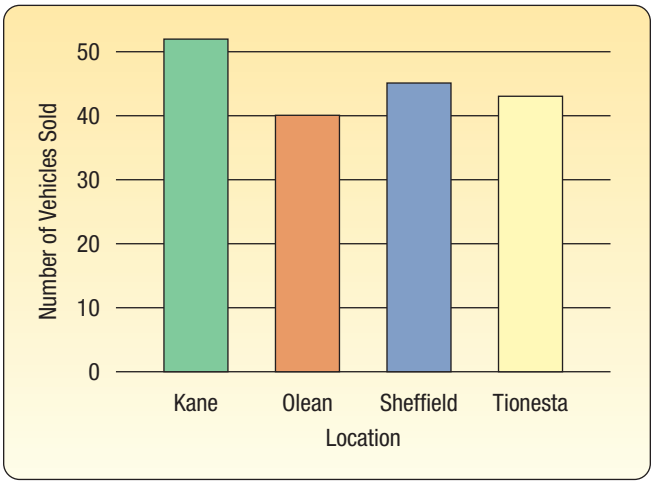


CHART 2–1 Number of Vehicles Sold by Location

Another useful type of chart for depicting qualitative information is a **pie chart.**

PIE CHART A chart that shows the proportion or percentage that each class represents of the total number of frequencies.

We explain the details of constructing a pie chart using the information in Table 2–3, which shows the frequency and percent of cars sold by the Applewood Auto Group for each vehicle type.

TABLE 2–3 Vehicle Sales by Type at Applewood Auto Group

Vehicle Type	Number Sold	Percent Sold
Sedan	72	40
SUV	54	30
Compact	27	15
Truck	18	10
Hybrid	9	5
Total	180	100

The first step to develop a pie chart is to mark the percentages 0, 5, 10, 15, and so on evenly around the circumference of a circle (see Chart 2–2). To plot the 40% of total sales represented by sedans, draw a line from the center of the circle to 0 and another line from the center of the circle to 40%. The area in this "slice" represents the number of sedans sold as a percentage of the total sales. Next, add the SUV's percentage of total sales, 30%, to the sedan's percentage of total sales, 40%. The result is 70%. Draw a line from the center of the circle to 70%, so the area between 40 and 70 shows the sales of SUVs as a percentage of total sales. Continuing, add the 15% of total sales for compact vehicles, which gives us a total of 85%. Draw a line from the center of the circle to 85, so the "slice" between 70% and 85% represents the number of compact vehicles sold as a percentage of the total sales. The remaining 10% for truck sales and 5% for hybrid sales are added to the chart using the same method.

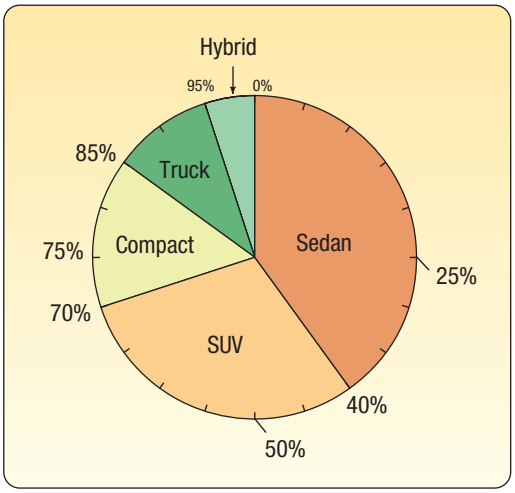
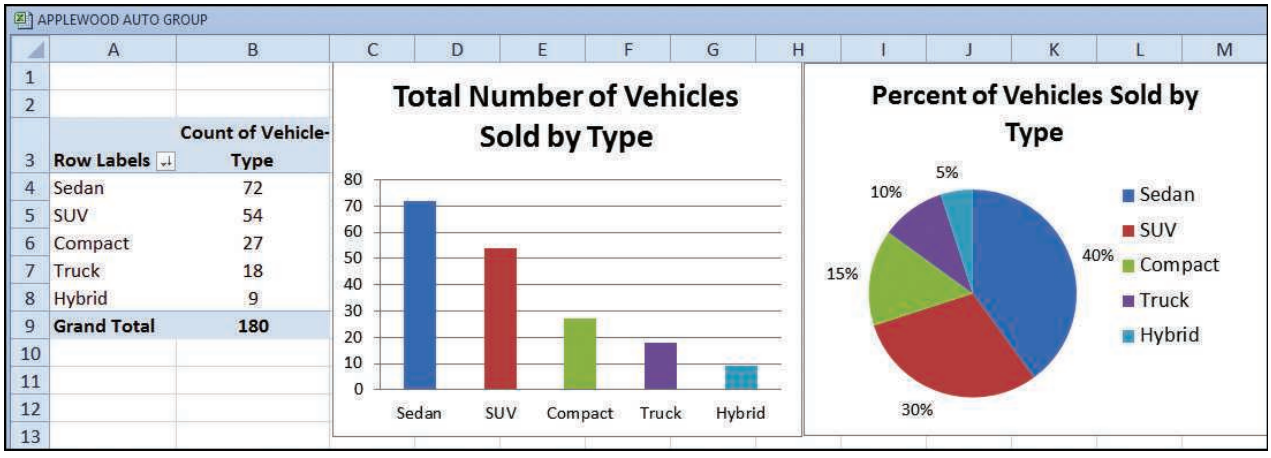


CHART 2–2 Pie Chart of Vehicles by Type

Because each slice of the pie represents the relative frequency of each vehicle type as a percentage of the total sales, we can easily compare them:

- The largest percentage of sales is for sedans.
- Sedans and SUVs together account for 70% of vehicle sales.
- Hybrids account for 5% of vehicle sales, in spite of being on the market for only a few years.

We can use Excel software to quickly count the number of cars for each vehicle type and create the frequency table, bar chart, and pie chart shown in the following summary. The Excel tool is called a Pivot Table. The instructions to produce these descriptive statistics and charts are given in Appendix C.



Pie and bar charts both serve to illustrate frequency and relative frequency tables. When is a pie chart preferred to a bar chart? In most cases, pie charts are used to show and compare the relative differences in the percentage of observations for each value or class of a qualitative variable. Bar charts are preferred when the goal is to compare the number or frequency of observations for each value or class of a qualitative variable. The following Example/Solution shows another application of bar and pie charts.

EXAMPLE

SkiLodges.com is test marketing its new website and is interested in how easy its website design is to navigate. It randomly selected 200 regular Internet users and asked them to perform a search task on the website. Each person was asked to rate the relative ease of navigation as poor, good, excellent, or awesome. The results are shown in the following table:

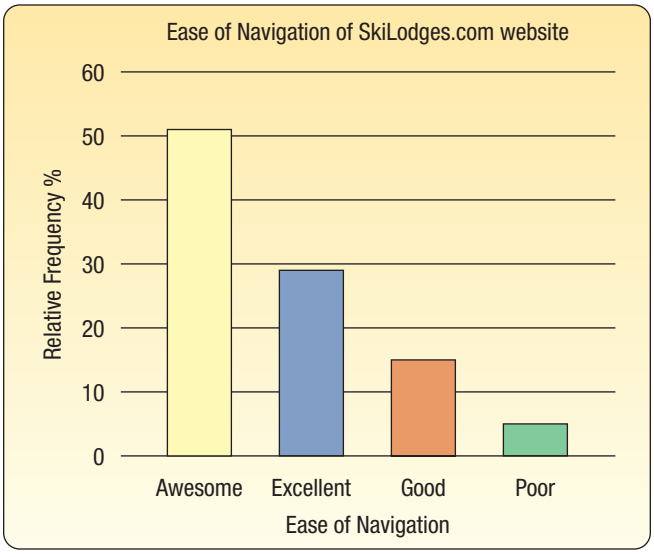
Awesome	102
Excellent	58
Good	30
Poor	10

1. What type of measurement scale is used for ease of navigation?
2. Draw a bar chart for the survey results.
3. Draw a pie chart for the survey results.

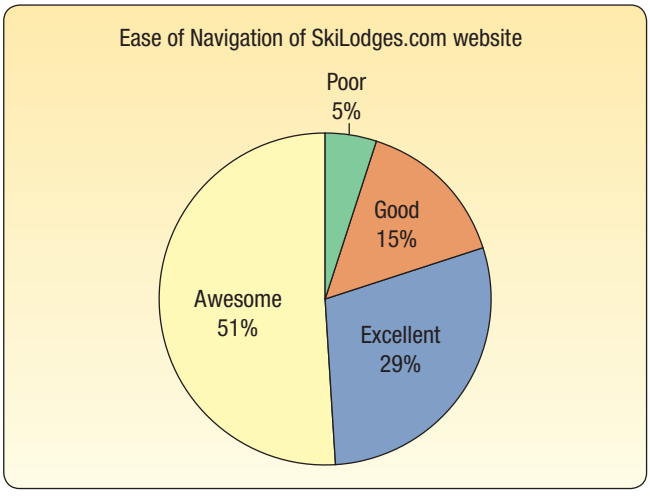
SOLUTION

The data are measured on an ordinal scale. That is, the scale is ranked in relative ease of navigation when moving from “awesome” to “poor.” The interval between each rating is unknown so it is impossible, for example, to conclude that a rating of good is twice the value of a poor rating.

We can use a bar chart to graph the data. The vertical scale shows the relative frequency and the horizontal scale shows the values of the ease-of-navigation variable.



A pie chart can also be used to graph these data. The pie chart emphasizes that more than half of the respondents rate the relative ease of using the website awesome.



SELF-REVIEW 2–1

The answers are in Appendix E.

DeCenzo Specialty Food and Beverage Company has been serving a cola drink with an additional flavoring, Cola-Plus, that is very popular among its customers. The company is interested in customer preferences for Cola-Plus versus Coca-Cola, Pepsi, and a lemon-lime beverage. They ask 100 randomly sampled customers to take a taste test and select the beverage they prefer most. The results are shown in the following table:

Beverage	Number
Cola-Plus	40
Coca-Cola	25
Pepsi	20
Lemon-Lime	15
Total	100

(a) Is the data qualitative or quantitative? Why?
(b) What is the table called? What does it show?
(c) Develop a bar chart to depict the information.
(d) Develop a pie chart using the relative frequencies.

EXERCISES

The answers to the odd-numbered exercises are at the end of the book in Appendix D.

1. A pie chart shows the relative market share of cola products. The "slice" for Pepsi-Cola has a central angle of 90 degrees. What is its market share?

2. In a marketing study, 100 consumers were asked to select the best digital music player from the iPod, the iRiver, and the Magic Star MP3. To summarize the consumer responses with a frequency table, how many classes would the frequency table have?

3. A total of 1,000 residents in Minnesota were asked which season they preferred. One hundred liked winter best, 300 liked spring, 400 liked summer, and 200 liked fall. Develop a frequency table and a relative frequency table to summarize this information.

4. Two thousand frequent business travelers are asked which midwestern city they prefer: Indianapolis, Saint Louis, Chicago, or Milwaukee. One hundred liked Indianapolis best, 450 liked Saint Louis, 1,300 liked Chicago, and the remainder preferred Milwaukee. Develop a frequency table and a relative frequency table to summarize this information.

5. Wellstone Inc. produces and markets replacement covers for cell phones in five different colors: bright white, metallic black, magnetic lime, tangerine orange, and fusion red. To estimate the demand for each color, the company set up a kiosk in the Mall of America for several hours and asked randomly selected people which cover color was their favorite. The results follow:

Bright white	130
Metallic black	104
Magnetic lime	325
Tangerine orange	455
Fusion red	286

a. What is the table called?
b. Draw a bar chart for the table.
c. Draw a pie chart.
d. If Wellstone Inc. plans to produce 1 million cell phone covers, how many of each color should it produce?

6. A small business consultant is investigating the performance of several companies. The fourth-quarter sales for last year (in thousands of dollars) for the selected companies were:

Company	Fourth-Quarter Sales ($ thousands)
Hoden Building Products	$ 1,645.2
J & R Printing Inc.	4,757.0
Long Bay Concrete Construction	8,913.0
Mancell Electric and Plumbing	627.1
Maxwell Heating and Air Conditioning	24,612.0
Mizelle Roofing & Sheet Metals	191.9

The consultant wants to include a chart in his report comparing the sales of the six companies. Use a bar chart to compare the fourth-quarter sales of these corporations and write a brief report summarizing the bar chart.

LO2-3
Summarize quantitative variables with frequency and relative frequency distributions.

CONSTRUCTING FREQUENCY DISTRIBUTIONS

In Chapter 1 and earlier in this chapter, we distinguished between qualitative and quantitative data. In the previous section, using the Applewood Automotive Group data, we summarized two qualitative variables: the location of the sale and the type of vehicle sold. We created frequency and relative frequency tables and depicted the results in bar and pie charts.

The Applewood Auto Group data also includes several quantitative variables: the age of the buyer, the profit earned on the sale of the vehicle, and the number of previous purchases. Suppose Ms. Ball wants to summarize last month's sales by profit earned for each vehicle. We can describe profit using a **frequency distribution.**

FREQUENCY DISTRIBUTION A grouping of quantitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

How do we develop a frequency distribution? The following example shows the steps to construct a frequency distribution. Remember, our goal is to construct tables, charts, and graphs that will quickly summarize the data by showing the location, extreme values, and shape of the data's distribution.

EXAMPLE

Ms. Kathryn Ball of the Applewood Auto Group wants to summarize the quantitative variable profit with a frequency distribution and display the distribution with charts and graphs. With this information, Ms. Ball can easily answer the following questions: What is the typical profit on each sale? What is the largest or maximum profit on any sale? What is the smallest or minimum profit on any sale? Around what value do the profits tend to cluster?

SOLUTION

To begin, we need the profits for each of the 180 vehicle sales listed in Table 2–4. This information is called raw or ungrouped data because it is simply a listing

TABLE 2–4 Profit on Vehicles Sold Last Month by the Applewood Auto Group

$1,387	$2,148	$2,201	$ 963	$ 820	$2,230	$3,043	$2,584	$2,370
1,754	2,207	996	1,298	1,266	2,341	1,059	2,666	2,637
1,817	2,252	2,813	1,410	1,741	3,292	1,674	2,991	1,426
1,040	1,428	323	1,553	1,772	1,108	1,807	934	2,944
1,273	1,889	352	1,648	1,932	1,295	2,056	2,063	2,147
1,529	1,166	482	2,071	2,350	1,344	2,236	2,083	1,973
3,082	1,320	1,144	2,116	2,422	1,906	2,928	2,856	2,502
1,951	2,265	1,485	1,500	2,446	1,952	1,269	2,989	783
2,692	1,323	1,509	1,549	369	2,070	1,717	910	1,538
1,206	1,760	1,638	2,348	978	2,454	1,797	1,536	2,339
1,342	1,919	1,961	2,498	1,238	1,606	1,955	1,957	2,700
443	2,357	2,127	294	1,818	1,680	2,199	2,240	2,222
754	2,866	2,430	1,115	1,824	1,827	2,482	2,695	2,597
1,621	732	1,704	1,124	1,907	1,915	2,701	1,325	2,742
870	1,464	1,876	1,532	1,938	2,084	3,210	2,250	1,837
1,174	1,626	2,010	1,688	1,940	2,639	377	2,279	2,842
1,412	1,762	2,165	1,822	2,197	842	1,220	2,626	2,434
1,809	1,915	2,231	1,897	2,646	1,963	1,401	1,501	1,640
2,415	2,119	2,389	2,445	1,461	2,059	2,175	1,752	1,821
1,546	1,766	335	2,886	1,731	2,338	1,118	2,058	2,487

Maximum

Minimum

of the individual, observed profits. It is possible to search the list and find the smallest or minimum profit ($294) and the largest or maximum profit ($3,292), but that is about all. It is difficult to determine a typical profit or to visualize where the profits tend to cluster. The raw data are more easily interpreted if we summarize the data with a frequency distribution. The steps to create this frequency distribution follow.

Step 1: **Decide on the number of classes.** A useful recipe to determine the number of classes (k) is the "2 to the k rule." This guide suggests you select the smallest number (k) for the number of classes such that 2^k (in words, 2 raised to the power of k) is greater than the number of observations (n). In the Applewood Auto Group example, there were 180 vehicles sold. So $n = 180$. If we try $k = 7$, which means we would use 7 classes, $2^7 = 128$, which is less than 180. Hence, 7 is too few classes. If we let $k = 8$, then $2^8 = 256$, which is greater than 180. So the recommended number of classes is 8.

Step 2: **Determine the class interval.** Generally, the **class interval** is the same for all classes. The classes all taken together must cover at least the distance from the minimum value in the data up to the maximum value. Expressing these words in a formula:

$$i \geq \frac{\textit{Maximum Value} - \textit{Minimum Value}}{k}$$

where i is the class interval, and k is the number of classes.

For the Applewood Auto Group, the minimum value is $294 and the maximum value is $3,292. If we need 8 classes, the interval should be:

$$i \geq \frac{\textit{Maximum Value} - \textit{Minimum Value}}{k} = \frac{\$3{,}292 - \$294}{8} = \$374.75$$

In practice, this interval size is usually rounded up to some convenient number, such as a multiple of 10 or 100. The value of $400 is a reasonable choice.

Step 3: **Set the individual class limits.** State clear class limits so you can put each observation into only one category. This means you must avoid overlapping or unclear class limits. For example, classes such as "$1,300–$1,400" and "$1,400–$1,500" should not be used because it is not clear whether the value $1,400 is in the first or second class. In this text, we will generally use the format $1,300 **up to** $1,400 and $1,400 **up to** $1,500 and so on. With this format, it is clear that $1,399 goes into the first class and $1,400 in the second.

Because we always round the class interval up to get a convenient class size, we cover a larger than necessary range. For example, using 8 classes with an interval of $400 in the Applewood Auto Group example results in a range of 8($400) = $3,200. The actual range is $2,998, found by ($3,292 − $294). Comparing that value to $3,200, we have an excess of $202. Because we need to cover only the range (*Maximum* − *Minimum*), it is natural to put approximately equal amounts of the excess in each of the two tails. Of course, we also should select convenient class limits. A guideline is to make the lower limit of the first class a multiple of the class interval. Sometimes this is not possible, but the lower limit should at least be rounded. So here are the classes we could use for these data.

Classes
$ 200 up to $ 600
600 up to 1,000
1,000 up to 1,400
1,400 up to 1,800
1,800 up to 2,200
2,200 up to 2,600
2,600 up to 3,000
3,000 up to 3,400

Step 4: **Tally the vehicle profit into the classes and determine the number of observations in each class.** To begin, the profit from the sale of the first vehicle in Table 2–4 is $1,387. It is tallied in the $1,000 up to $1,400 class. The second profit in the first row of Table 2–4 is $2,148. It is tallied in the $1,800 up to $2,200 class. The other profits are tallied in a similar manner. When all the profits are tallied, the table would appear as:

Profit	Frequency
$ 200 up to $ 600	𝍸 |||
600 up to 1,000	𝍸 𝍸 |
1,000 up to 1,400	𝍸 𝍸 𝍸 𝍸 |||
1,400 up to 1,800	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||
1,800 up to 2,200	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸
2,200 up to 2,600	𝍸 𝍸 𝍸 𝍸 𝍸 ||
2,600 up to 3,000	𝍸 𝍸 𝍸 ||||
3,000 up to 3,400	||||

The number of observations in each class is called the **class frequency.** In the $200 up to $600 class there are 8 observations, and in the $600 up to $1,000 class there are 11 observations. Therefore, the class frequency in the first class is 8 and the class frequency in the second class is 11. There are a total of 180 observations in the entire set of data. So the sum of all the frequencies should be equal to 180. The results of the frequency distribution are in Table 2–5.

TABLE 2–5 Frequency Distribution of Profit for Vehicles Sold Last Month at Applewood Auto Group

Profit	Frequency
$ 200 up to $ 600	8
600 up to 1,000	11
1,000 up to 1,400	23
1,400 up to 1,800	38
1,800 up to 2,200	45
2,200 up to 2,600	32
2,600 up to 3,000	19
3,000 up to 3,400	4
Total	180

Now that we have organized the data into a frequency distribution (see Table 2–5), we can summarize the profits of the vehicles for the Applewood Auto Group. Observe the following:

1. The profits from vehicle sales range between $200 and $3,400.
2. The vehicle profits are classified using a class interval of $400. The class interval is determined by subtracting consecutive lower or upper class limits. For

example, the lower limit of the first class is $200, and the lower limit of the second class is $600. The difference is the class interval of $400.

3. The profits are concentrated between $1,000 and $3,000. The profit on 157 vehicles, or 87%, was within this range.
4. For each class, we can determine the typical profit or **class midpoint.** It is halfway between the lower or upper limits of two consecutive classes. It is computed by adding the lower or upper limits of consecutive classes and dividing by 2. Referring to Table 2–5, the lower class limit of the first class is $200, and the next class limit is $600. The class midpoint is $400, found by ($600 + $200)/2. The midpoint best represents, or is typical of, the profits of the vehicles in that class. Applewood sold 8 vehicles with a typical profit of $400.
5. The largest concentration, or highest frequency, of vehicles sold is in the $1,800 up to $2,200 class. There are 45 vehicles in this class. The class midpoint is $2,000. So we say that the typical profit in the class with the highest frequency is $2,000.

By presenting this information to Ms. Ball, we give her a clear picture of the distribution of the vehicle profits for last month.

We admit that arranging the information on profits into a frequency distribution does result in the loss of some detailed information. That is, by organizing the data into a frequency distribution, we cannot pinpoint the exact profit on any vehicle, such as $1,387, $2,148, or $2,201. Further, we cannot tell that the actual minimum profit for any vehicle sold is $294 or that the maximum profit was $3,292. However, the lower limit of the first class and the upper limit of the last class convey essentially the same meaning. Likely, Ms. Ball will make the same judgment if she knows the smallest profit is about $200 that she will if she knows the exact profit is $292. The advantages of summarizing the 180 profits into a more understandable and organized form more than offset this disadvantage.

When we summarize raw data with frequency distributions, equal class intervals are preferred. However, in certain situations unequal class intervals may be necessary to avoid a large number of classes with very small frequencies. Such is the case in Table 2–6. The U.S. Internal Revenue Service uses unequal-sized class intervals for adjusted gross income on individual tax returns to summarize the number of individual tax returns. If we use our method to find equal class intervals, the 2^k rule results in 25 classes, and

STATISTICS IN ACTION

In 1788, James Madison, John Jay, and Alexander Hamilton anonymously published a series of essays entitled *The Federalist.* These Federalist papers were an attempt to convince the people of New York that they should ratify the Constitution. In the course of history, the authorship of most of these papers became known, but 12 remained contested. Through the use of statistical analysis, and particularly studying the frequency distributions of various words, we can now conclude that James Madison is the likely author of the 12 papers. In fact, the statistical evidence that Madison is the author is overwhelming.

TABLE 2–6 Adjusted Gross Income for Individuals Filing Income Tax Returns

Adjusted Gross Income	Number of Returns (in thousands)
No adjusted gross income	178.2
$ 1 up to 5,000	1,204.6
5,000 up to 10,000	2,595.5
10,000 up to 15,000	3,142.0
15,000 up to 20,000	3,191.7
20,000 up to 25,000	2,501.4
25,000 up to 30,000	1,901.6
30,000 up to 40,000	2,502.3
40,000 up to 50,000	1,426.8
50,000 up to 75,000	1,476.3
75,000 up to 100,000	338.8
100,000 up to 200,000	223.3
200,000 up to 500,000	55.2
500,000 up to 1,000,000	12.0
1,000,000 up to 2,000,000	5.1
2,000,000 up to 10,000,000	3.4
10,000,000 or more	0.6

a class interval of $400,000, assuming $0 and $10,000,000 as the minimum and maximum values for adjusted gross income. Using equal class intervals, the first 13 classes in Table 2–6 would be combined into one class of about 99.9% of all tax returns and 24 classes for the 0.1% of the returns with an adjusted gross income above $400,000. Using equal class intervals does not provide a good understanding of the raw data. In this case, good judgment in the use of unequal class intervals, as demonstrated in Table 2–6, is required to show the distribution of the number of tax returns filed, especially for incomes under $500,000.

SELF-REVIEW 2–2

In the first quarter of last year, the 11 members of the sales staff at Master Chemical Company earned the following commissions:

$1,650 $1,475 $1,510 $1,670 $1,595 $1,760 $1,540 $1,495 $1,590 $1,625 $1,510

(a) What are the values such as $1,650 and $1,475 called?
(b) Using $1,400 up to $1,500 as the first class, $1,500 up to $1,600 as the second class, and so forth, organize the quarterly commissions into a frequency distribution.
(c) What are the numbers in the right column of your frequency distribution called?
(d) Describe the distribution of quarterly commissions, based on the frequency distribution. What is the largest concentration of commissions earned? What is the smallest, and the largest? What is the typical amount earned?

Relative Frequency Distribution

It may be desirable, as we did earlier with qualitative data, to convert class frequencies to relative class frequencies to show the proportion of the total number of observations in each class. In our vehicle profits, we may want to know what percentage of the vehicle profits are in the $1,000 up to $1,400 class. To convert a frequency distribution to a *relative* frequency distribution, each of the class frequencies is divided by the total number of observations. From the distribution of vehicle profits, Table 2–5, the relative frequency for the $1,000 up to $1,400 class is 0.128, found by dividing 23 by 180. That is, profit on 12.8% of the vehicles sold is between $1,000 and $1,400. The relative frequencies for the remaining classes are shown in Table 2–7.

TABLE 2–7 Relative Frequency Distribution of Profit for Vehicles Sold Last Month at Applewood Auto Group

Profit	Frequency	Relative Frequency	Found by
$ 200 up to $ 600	8	.044	8/180
600 up to 1,000	11	.061	11/180
1,000 up to 1,400	23	.128	23/180
1,400 up to 1,800	38	.211	38/180
1,800 up to 2,200	45	.250	45/180
2,200 up to 2,600	32	.178	32/180
2,600 up to 3,000	19	.106	19/180
3,000 up to 3,400	4	.022	4/180
Total	180	1.000	

APPLEWOOD AUTO GROUP

	A	B	C
1	Profit Class	Frequency	Relative Frequency
2	200-600	8	4.44%
3	600-1000	11	6.11%
4	1000-1400	23	12.78%
5	1400-1800	38	21.11%
6	1800-2200	45	25.00%
7	2200-2600	32	17.78%
8	2600-3000	19	10.56%
9	3000-3400	4	2.22%
10	Grand Total	180	100.00%

There are many software packages that perform statistical calculations. Throughout this text, we will show the output from Microsoft Excel, MegaStat (a Microsoft Excel add-in), and Minitab (a statistical software package). Because Excel is most readily available, it is used most frequently.

Within the earlier Graphic Presentation of Qualitative Data section, we used the Pivot Table tool in Excel to create a frequency table. To create the table to the left, we use the same Excel tool to

compute frequency and relative frequency distributions for the profit variable in the Applewood Auto Group data. The necessary steps are given in the Software Commands section in Appendix C.

SELF-REVIEW 2–3

Barry Bonds of the San Francisco Giants established a new single-season Major League Baseball home run record by hitting 73 home runs during the 2001 season. Listed below is the sorted distance of each of the 73 home runs.

320	320	347	350	360	360	360	361	365	370
370	375	375	375	375	380	380	380	380	380
380	390	390	391	394	396	400	400	400	400
405	410	410	410	410	410	410	410	410	410
410	410	411	415	415	416	417	417	420	420
420	420	420	420	420	420	429	430	430	430
430	430	435	435	436	440	440	440	440	440
450	480	488							

(a) For this data, show that seven classes would be used to create a frequency distribution using the 2^k rule.
(b) Show that a class interval of 30 would summarize the data in seven classes.
(c) Construct frequency and relative frequency distributions for the data with seven classes and a class interval of 30. Start the first class with a lower limit of 300.
(d) How many home runs traveled a distance of 360 up to 390 feet?
(e) What percentage of the home runs traveled a distance of 360 up to 390 feet?
(f) What percentage of the home runs traveled a distance of 390 feet or more?

EXERCISES

This FILE icon indicates that the data are available at the text website: www.mhhe.com/Lind17e. You will be able to download the data directly into Excel or Minitab from this site.

7. A set of data consists of 38 observations. How many classes would you recommend for the frequency distribution?
8. A set of data consists of 45 observations between $0 and $29. What size would you recommend for the class interval?
9. A set of data consists of 230 observations between $235 and $567. What class interval would you recommend?
10. A set of data contains 53 observations. The minimum value is 42 and the maximum value is 129. The data are to be organized into a frequency distribution.
   a. How many classes would you suggest?
   b. What would you suggest as the lower limit of the first class?
11. FILE Wachesaw Manufacturing Inc. produced the following number of units in the last 16 days.

27	27	27	28	27	25	25	28
26	28	26	28	31	30	26	26

The information is to be organized into a frequency distribution.
   a. How many classes would you recommend?
   b. What class interval would you suggest?
   c. What lower limit would you recommend for the first class?
   d. Organize the information into a frequency distribution and determine the relative frequency distribution.
   e. Comment on the shape of the distribution.

12. FILE The Quick Change Oil Company has a number of outlets in the metropolitan Seattle area. The daily number of oil changes at the Oak Street outlet in the past 20 days are:

65	98	55	62	79	59	51	90	72	56
70	62	66	80	94	79	63	73	71	85

The data are to be organized into a frequency distribution.

a. How many classes would you recommend?
b. What class interval would you suggest?
c. What lower limit would you recommend for the first class?
d. Organize the number of oil changes into a frequency distribution.
e. Comment on the shape of the frequency distribution. Also determine the relative frequency distribution.

13. FILE The manager of the BiLo Supermarket in Mt. Pleasant, Rhode Island, gathered the following information on the number of times a customer visits the store during a month. The responses of 51 customers were:

5	3	3	1	4	4	5	6	4	2	6	6	6	7	1
1	14	1	2	4	4	4	5	6	3	5	3	4	5	6
8	4	7	6	5	9	11	3	12	4	7	6	5	15	1
1	10	8	9	2	12									

a. Starting with 0 as the lower limit of the first class and using a class interval of 3, organize the data into a frequency distribution.
b. Describe the distribution. Where do the data tend to cluster?
c. Convert the distribution to a relative frequency distribution.

14. FILE The food services division of Cedar River Amusement Park Inc. is studying the amount of money spent per day on food and drink by families who visit the amusement park. A sample of 40 families who visited the park yesterday revealed they spent the following amounts:

$77	$18	$63	$84	$38	$54	$50	$59	$54	$56	$36	$26	$50	$34	$44
41	58	58	53	51	62	43	52	53	63	62	62	65	61	52
60	60	45	66	83	71	63	58	61	71					

a. Organize the data into a frequency distribution, using seven classes and 15 as the lower limit of the first class. What class interval did you select?
b. Where do the data tend to cluster?
c. Describe the distribution.
d. Determine the relative frequency distribution.

LO2-4
Display a distribution using a histogram or frequency polygon.

GRAPHIC PRESENTATION OF A DISTRIBUTION

Sales managers, stock analysts, hospital administrators, and other busy executives often need a quick picture of the distributions of sales, stock prices, or hospital costs. These distributions can often be depicted by the use of charts and graphs. Three charts that will help portray a frequency distribution graphically are the histogram, the frequency polygon, and the cumulative frequency polygon.

Histogram

A **histogram** for a frequency distribution based on quantitative data is similar to the bar chart showing the distribution of qualitative data. The classes are marked on the

horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars. However, there is one important difference based on the nature of the data. Quantitative data are usually measured using scales that are continuous, not discrete. Therefore, the horizontal axis represents all possible values, and the bars are drawn adjacent to each other to show the continuous nature of the data.

HISTOGRAM A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to each other.

EXAMPLE

Below is the frequency distribution of the profits on vehicle sales last month at the Applewood Auto Group.

Profit	Frequency
$ 200 up to $ 600	8
600 up to 1,000	11
1,000 up to 1,400	23
1,400 up to 1,800	38
1,800 up to 2,200	45
2,200 up to 2,600	32
2,600 up to 3,000	19
3,000 up to 3,400	4
Total	180

Construct a histogram. What observations can you reach based on the information presented in the histogram?

SOLUTION

The class frequencies are scaled along the vertical axis (*Y*-axis) and either the class limits or the class midpoints along the horizontal axis. To illustrate the construction of the histogram, the first three classes are shown in Chart 2–3.

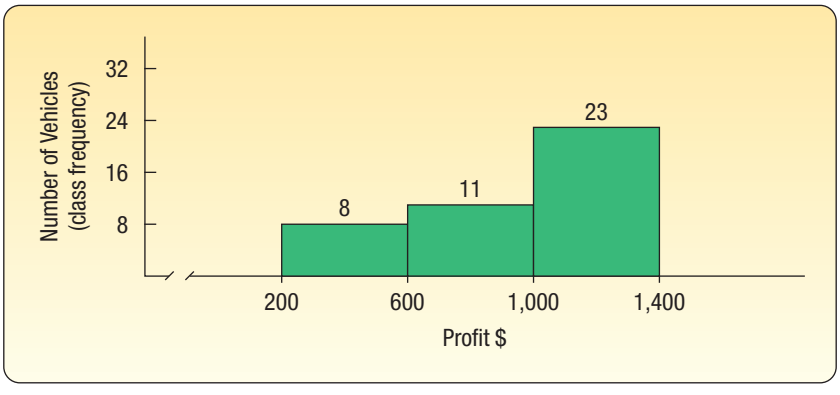


CHART 2–3 Construction of a Histogram

From Chart 2–3 we note the profit on eight vehicles was $200 up to $600. Therefore, the height of the column for that class is 8. There are 11 vehicle sales where the profit was $600 up to $1,000. So, logically, the height of that column is 11. The height of the bar represents the number of observations in the class.

This procedure is continued for all classes. The complete histogram is shown in Chart 2–4. Note that there is no space between the bars. This is a feature of the histogram. Why is this so? Because the variable profit, plotted on the horizontal axis, is a continuous variable. In a bar chart, the scale of measurement is usually nominal and the vertical bars are separated. This is an important distinction between the histogram and the bar chart.

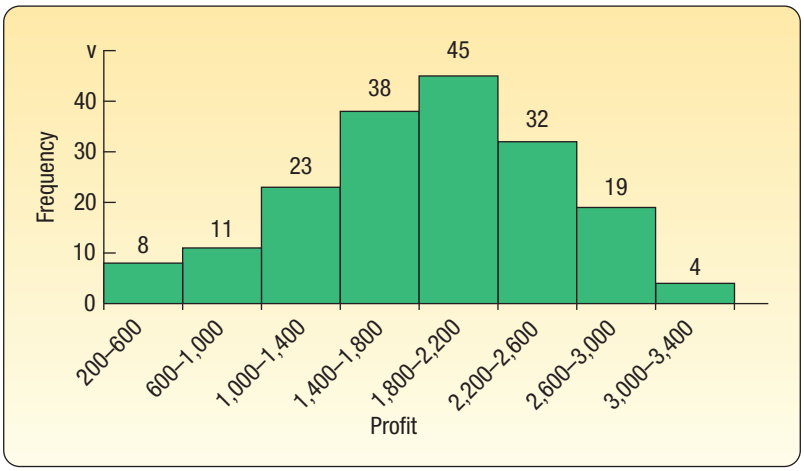


CHART 2–4 Histogram of the Profit on 180 Vehicles Sold at the Applewood Auto Group

We can make the following statements using Chart 2–4. They are the same as the observations based on Table 2–5.

1. The profits from vehicle sales range between $200 and $3,400.
2. The vehicle profits are classified using a class interval of $400. The class interval is determined by subtracting consecutive lower or upper class limits. For example, the lower limit of the first class is $200, and the lower limit of the second class is $600. The difference is the class interval or $400.
3. The profits are concentrated between $1,000 and $3,000. The profit on 157 vehicles, or 87%, was within this range.
4. For each class, we can determine the typical profit or class midpoint. It is halfway between the lower or upper limits of two consecutive classes. It is computed by adding the lower or upper limits of consecutive classes and dividing by 2. Referring to Chart 2–4, the lower class limit of the first class is $200, and the next class limit is $600. The class midpoint is $400, found by ($600 + $200)/2. The midpoint best represents, or is typical of, the profits of the vehicles in that class. Applewood sold 8 vehicles with a typical profit of $400.
5. The largest concentration, or highest frequency of vehicles sold, is in the $1,800 up to $2,200 class. There are 45 vehicles in this class. The class midpoint is $2,000. So we say that the typical profit in the class with the highest frequency is $2,000.

Thus, the histogram provides an easily interpreted visual representation of a frequency distribution. We should also point out that we would have made the same observations and the shape of the histogram would have been the same had we used a relative frequency distribution instead of the actual frequencies. That is, if we use the relative frequencies of Table 2–7, the result is a histogram of the same shape as Chart 2–4. The only difference is that the vertical axis would have been reported in percentage of vehicles instead of the number of vehicles. The Excel commands to create Chart 2–4 are given in Appendix C.

STATISTICS IN ACTION

Florence Nightingale is known as the founder of the nursing profession. However, she also saved many lives by using statistical analysis. When she encountered an unsanitary condition or an undersupplied hospital, she improved the conditions and then used statistical data to document the improvement. Thus, she was able to convince others of the need for medical reform, particularly in the area of sanitation. She developed original graphs to demonstrate that, during the Crimean War, more soldiers died from unsanitary conditions than were killed in combat.

Frequency Polygon

A **frequency polygon** also shows the shape of a distribution and is similar to a histogram. It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies. The construction of a frequency polygon is illustrated in Chart 2–5. We use the profits from the cars sold last month at the Applewood Auto Group. The midpoint of each class is scaled on the *X*-axis and the class frequencies on the *Y*-axis. Recall that the class midpoint is the value at the center of a class and represents the typical values in that class. The class frequency is the number of observations in a particular class. The profit earned on the vehicles sold last month by the Applewood Auto Group is repeated below.

Profit	Midpoint	Frequency
$ 200 up to $ 600	$ 400	8
600 up to 1,000	800	11
1,000 up to 1,400	1,200	23
1,400 up to 1,800	1,600	38
1,800 up to 2,200	2,000	45
2,200 up to 2,600	2,400	32
2,600 up to 3,000	2,800	19
3,000 up to 3,400	3,200	4
Total		180

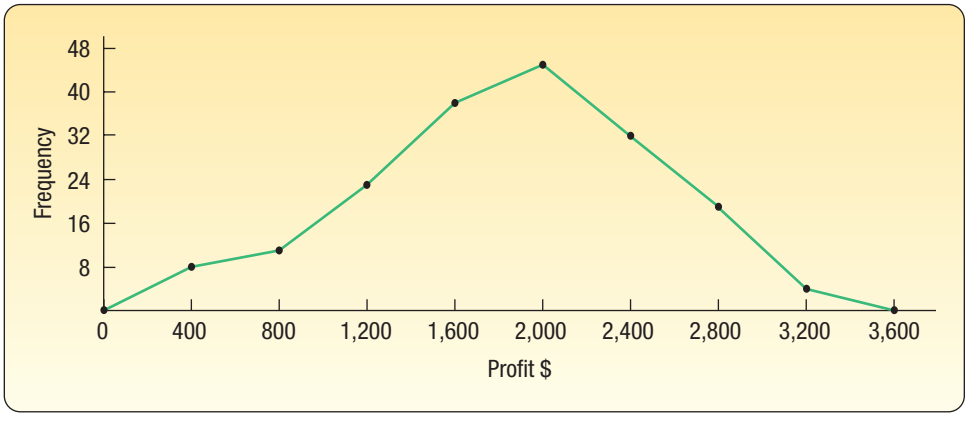


CHART 2–5 Frequency Polygon of Profit on 180 Vehicles Sold at Applewood Auto Group

As noted previously, the $200 up to $600 class is represented by the midpoint $400. To construct a frequency polygon, move horizontally on the graph to the midpoint, $400, and then vertically to 8, the class frequency, and place a dot. The *x* and the *y* values of this point are called the *coordinates.* The coordinates of the next point are $x = 800$ and $y = 11$. The process is continued for all classes. Then the points are connected in order. That is, the point representing the lowest class is joined to the one representing the second class and so on. Note in Chart 2–5 that, to complete the frequency polygon, midpoints of $0 and $3,600 are added to the *X*-axis to "anchor" the polygon at zero frequencies. These two values, $0 and $3,600, were derived by subtracting the class interval of $400 from the lowest midpoint ($400) and by adding $400 to the highest midpoint ($3,200) in the frequency distribution.

Both the histogram and the frequency polygon allow us to get a quick picture of the main characteristics of the data (highs, lows, points of concentration, etc.). Although the two representations are similar in purpose, the histogram has the advantage of depicting each class as a rectangle, with the height of the rectangular bar representing

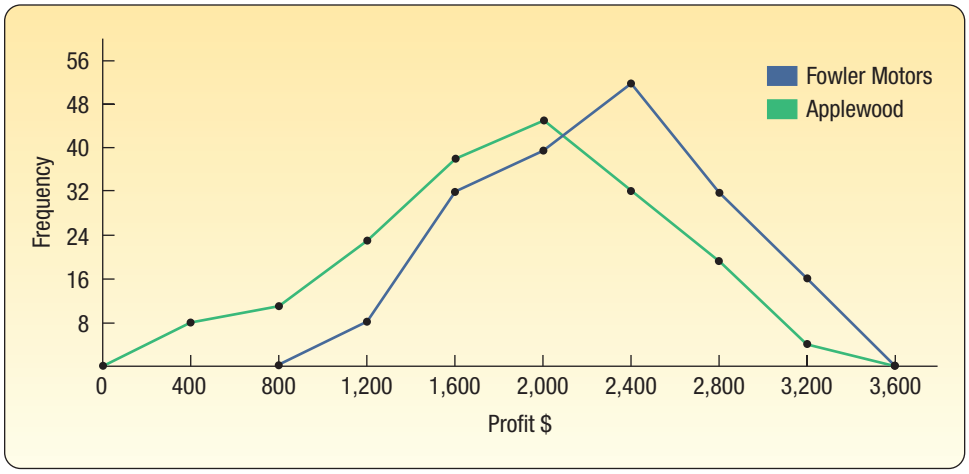


CHART 2–6 Distribution of Profit at Applewood Auto Group and Fowler Motors

the number in each class. The frequency polygon, in turn, has an advantage over the histogram. It allows us to compare directly two or more frequency distributions. Suppose Ms. Ball wants to compare the profit per vehicle sold at Applewood Auto Group with a similar auto group, Fowler Auto in Grayling, Michigan. To do this, two frequency polygons are constructed, one on top of the other, as in Chart 2–6. Two things are clear from the chart:

- The typical vehicle profit is larger at Fowler Motors—about $2,000 for Applewood and about $2,400 for Fowler.
- There is less variation or dispersion in the profits at Fowler Motors than at Applewood. The lower limit of the first class for Applewood is $0 and the upper limit is $3,600. For Fowler Motors, the lower limit is $800 and the upper limit is the same: $3,600.

The total number of cars sold at the two dealerships is about the same, so a direct comparison is possible. If the difference in the total number of cars sold is large, then converting the frequencies to relative frequencies and then plotting the two distributions would allow a clearer comparison.

SELF-REVIEW 2–4

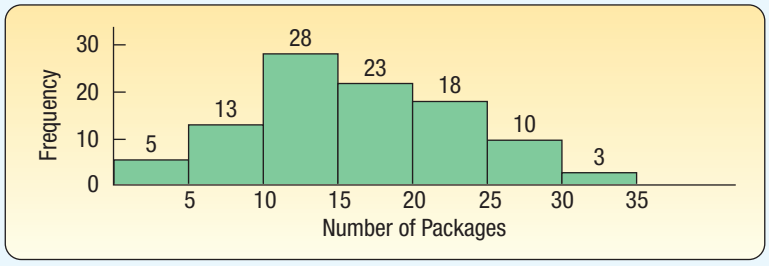
The annual imports of a selected group of electronic suppliers are shown in the following frequency distribution.

Imports ($ millions)	Number of Suppliers
2 up to 5	6
5 up to 8	13
8 up to 11	20
11 up to 14	10
14 up to 17	1

(a) Portray the imports as a histogram.
(b) Portray the imports as a relative frequency polygon.
(c) Summarize the important facets of the distribution (such as classes with the highest and lowest frequencies).

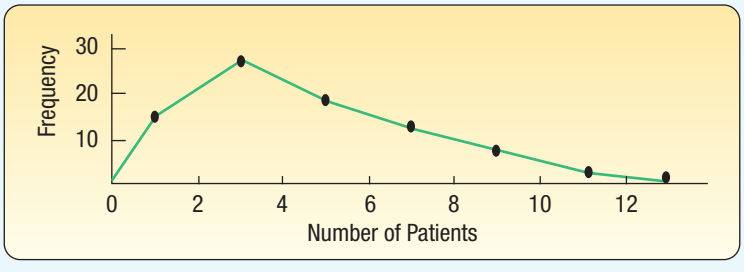
EXERCISES

15. Molly's Candle Shop has several retail stores in the coastal areas of North and South Carolina. Many of Molly's customers ask her to ship their purchases. The following chart shows the number of packages shipped per day for the last 100 days. For example, the first class shows that there were 5 days when the number of packages shipped was 0 up to 5.



a. What is this chart called?
b. What is the total number of packages shipped?
c. What is the class interval?
d. What is the number of packages shipped in the 10 up to 15 class?
e. What is the relative frequency of packages shipped in the 10 up to 15 class?
f. What is the midpoint of the 10 up to 15 class?
g. On how many days were there 25 or more packages shipped?

16. The following chart shows the number of patients admitted daily to Memorial Hospital through the emergency room.



a. What is the midpoint of the 2 up to 4 class?
b. How many days were 2 up to 4 patients admitted?
c. What is the class interval?
d. What is this chart called?

17. The following frequency distribution reports the number of frequent flier miles, reported in thousands, for employees of Brumley Statistical Consulting Inc. during the most recent quarter.

Frequent Flier Miles (000)	Number of Employees
0 up to 3	5
3 up to 6	12
6 up to 9	23
9 up to 12	8
12 up to 15	2
Total	50

a. How many employees were studied?
b. What is the midpoint of the first class?
c. Construct a histogram.
d. A frequency polygon is to be drawn. What are the coordinates of the plot for the first class?
e. Construct a frequency polygon.
f. Interpret the frequent flier miles accumulated using the two charts.

18. A large Internet retailer is studying the lead time (elapsed time between when an order is placed and when it is filled) for a sample of recent orders. The lead times are reported in days.

Lead Time (days)	Frequency
0 up to 5	6
5 up to 10	7
10 up to 15	12
15 up to 20	8
20 up to 25	7
Total	40

a. How many orders were studied?
b. What is the midpoint of the first class?
c. What are the coordinates of the first class for a frequency polygon?
d. Draw a histogram.
e. Draw a frequency polygon.
f. Interpret the lead times using the two charts.

Cumulative Distributions

Consider once again the distribution of the profits on vehicles sold by the Applewood Auto Group. Suppose we were interested in the number of vehicles that sold for a profit of less than $1,400. These values can be approximated by developing a **cumulative frequency distribution** and portraying it graphically in a **cumulative frequency polygon.** Or, suppose we were interested in the profit earned on the lowest-selling 40% of the vehicles. These values can be approximated by developing a **cumulative relative frequency distribution** and portraying it graphically in a **cumulative relative frequency polygon.**

EXAMPLE

The frequency distribution of the profits earned at Applewood Auto Group is repeated from Table 2–5.

Profit	Frequency
$ 200 up to $ 600	8
600 up to 1,000	11
1,000 up to 1,400	23
1,400 up to 1,800	38
1,800 up to 2,200	45
2,200 up to 2,600	32
2,600 up to 3,000	19
3,000 up to 3,400	4
Total	180

Construct a cumulative frequency polygon to answer the following question: sixty of the vehicles earned a profit of less than what amount? Construct a cumulative relative frequency polygon to answer this question: seventy-five percent of the vehicles sold earned a profit of less than what amount?

SOLUTION

As the names imply, a cumulative frequency distribution and a cumulative frequency polygon require *cumulative frequencies.* To construct a cumulative frequency distribution, refer to the preceding table and note that there were eight vehicles in which the profit earned was less than $600. Those 8 vehicles, plus the 11 in the next higher class, for a total of 19, earned a profit of less than $1,000. The cumulative frequency for the next higher class is 42, found by 8 + 11 + 23. This process is continued for all the classes. All the vehicles earned a profit of less than $3,400. (See Table 2–8.)

TABLE 2–8 Cumulative Frequency Distribution for Profit on Vehicles Sold Last Month at Applewood Auto Group

Profit	Cumulative Frequency	Found by
Less than $ 600	8	8
Less than 1,000	19	8 + 11
Less than 1,400	42	8 + 11 + 23
Less than 1,800	80	8 + 11 + 23 + 38
Less than 2,200	125	8 + 11 + 23 + 38 + 45
Less than 2,600	157	8 + 11 + 23 + 38 + 45 + 32
Less than 3,000	176	8 + 11 + 23 + 38 + 45 + 32 + 19
Less than 3,400	180	8 + 11 + 23 + 38 + 45 + 32 + 19 + 4

To construct a cumulative relative frequency distribution, we divide the cumulative frequencies by the total number of observations, 180. As shown in Table 2-9, the cumulative relative frequency of the fourth class is 80/180 = 44%. This means that 44% of the vehicles sold for less than $1,800.

TABLE 2–9 Cumulative Relative Frequency Distribution for Profit on Vehicles Sold Last Month at Applewood Auto Group

Profit	Cumulative Frequency	Cumulative Relative Frequency
Less than $ 600	8	8/180 = 0.044 = 4.4%
Less than $1,000	19	19/180 = 0.106 = 10.6%
Less than $1,400	42	42/180 = 0.233 = 23.3%
Less than $1,800	80	80/180 = 0.444 = 44.4%
Less than $2,200	125	125/180 = 0.694 = 69.4%
Less than $2,600	157	157/180 = 0.872 = 87.2%
Less than $3,000	176	176/180 = 0.978 = 97.8%
Less than $3,400	180	180/180 = 1.000 = 100%

To plot a cumulative frequency distribution, scale the upper limit of each class along the *X*-axis and the corresponding cumulative frequencies along the *Y*-axis. To provide additional information, you can label the vertical axis on the right in terms of cumulative relative frequencies. In the Applewood Auto Group,

the vertical axis on the left is labeled from 0 to 180 and on the right from 0 to 100%. Note, as an example, that 50% on the right axis should be opposite 90 vehicles on the left axis.

To begin, the first plot is at $x = 200$ and $y = 0$. None of the vehicles sold for a profit of less than $200. The profit on 8 vehicles was less than $600, so the next plot is at $x = 600$ and $y = 8$. Continuing, the next plot is $x = 1{,}000$ and $y = 19$. There were 19 vehicles that sold for a profit of less than $1,000. The rest of the points are plotted and then the dots connected to form Chart 2–7.

We should point out that the shape of the distribution is the same if we use cumulative relative frequencies instead of the cumulative frequencies. The only difference is that the vertical axis is scaled in percentages. In the following charts, a percentage scale is added to the right side of the graphs to help answer questions about cumulative relative frequencies.

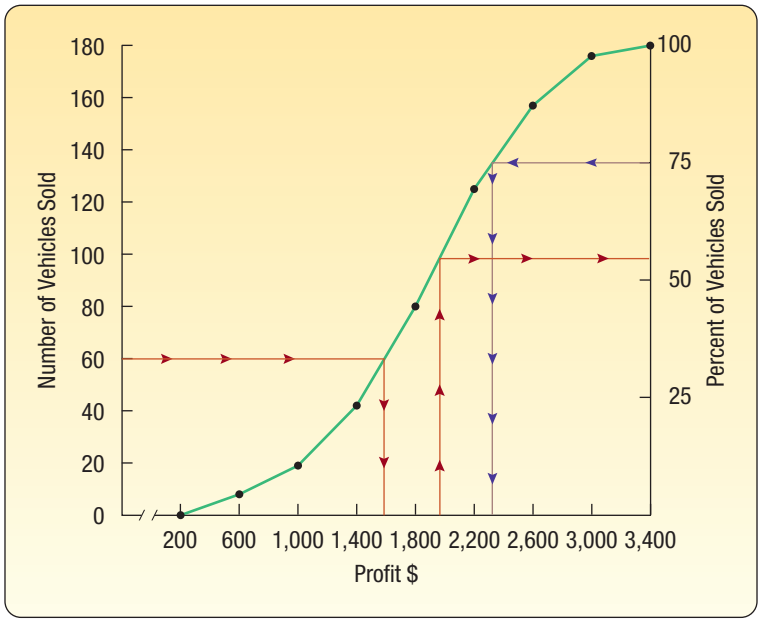


CHART 2–7 Cumulative Frequency Polygon for Profit on Vehicles Sold Last Month at Applewood Auto Group

Using Chart 2–7 to find the amount of profit on 75% of the cars sold, draw a horizontal line from the 75% mark on the right-hand vertical axis over to the polygon, then drop down to the *X*-axis and read the amount of profit. The value on the *X*-axis is about $2,300, so we estimate that 75% of the vehicles sold earned a profit of $2,300 or less for the Applewood group.

To find the highest profit earned on 60 of the 180 vehicles, we use Chart 2–7 to locate the value of 60 on the left-hand vertical axis. Next, we draw a horizontal line from the value of 60 to the polygon and then drop down to the *X*-axis and read the profit. It is about $1,600, so we estimate that 60 of the vehicles sold for a profit of less than $1,600. We can also make estimates of the percentage of vehicles that sold for less than a particular amount. To explain, suppose we want to estimate the percentage of vehicles that sold for a profit of less than $2,000. We begin by locating the value of $2,000 on the *X*-axis, move vertically to the polygon, and then horizontally to the vertical axis on the right. The value is about 56%, so we conclude 56% of the vehicles sold for a profit of less than $2,000.

SELF-REVIEW 2–5

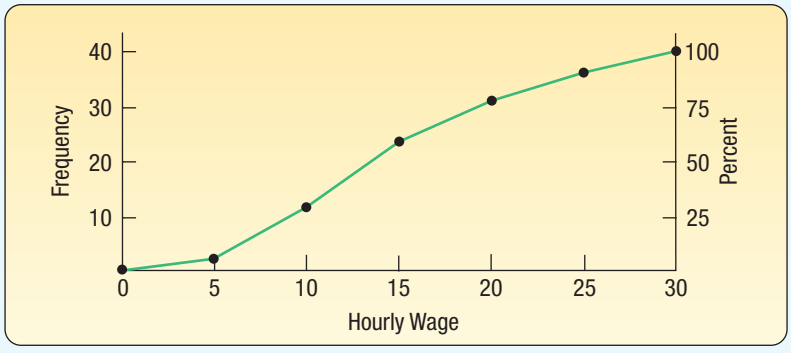
A sample of the hourly wages of 15 employees at Home Depot in Brunswick, Georgia, was organized into the following table.

Hourly Wages	Number of Employees
$ 8 up to $10	3
10 up to 12	7
12 up to 14	4
14 up to 16	1

(a) What is the table called?
(b) Develop a cumulative frequency distribution and portray the distribution in a cumulative frequency polygon.
(c) On the basis of the cumulative frequency polygon, how many employees earn less than $11 per hour?

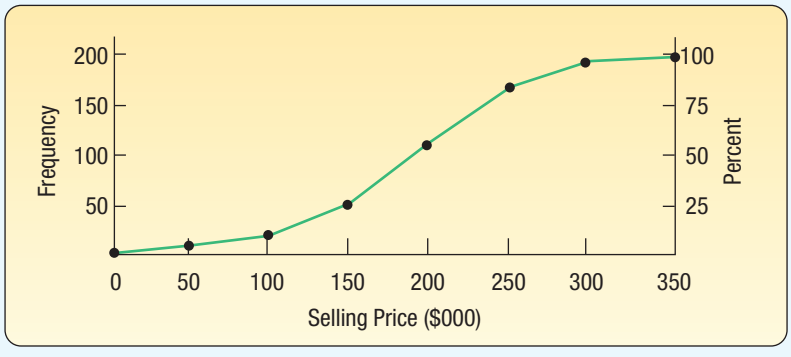
EXERCISES

19. The following cumulative frequency and the cumulative relative frequency polygon for the distribution of hourly wages of a sample of certified welders in the Atlanta, Georgia, area is shown in the graph.



a. How many welders were studied?
b. What is the class interval?
c. About how many welders earn less than $10.00 per hour?
d. About 75% of the welders make less than what amount?
e. Ten of the welders studied made less than what amount?
f. What percent of the welders make less than $20.00 per hour?

20. The cumulative frequency and the cumulative relative frequency polygon for a distribution of selling prices ($000) of houses sold in the Billings, Montana, area is shown in the graph.



a. How many homes were studied?
b. What is the class interval?
c. One hundred homes sold for less than what amount?
d. About 75% of the homes sold for less than what amount?
e. Estimate the number of homes in the $150,000 up to $200,000 class.
f. About how many homes sold for less than $225,000?

21. The frequency distribution representing the number of frequent flier miles accumulated by employees at Brumley Statistical Consulting Inc. is repeated from Exercise 17.

Frequent Flier Miles (000)	Frequency
0 up to 3	5
3 up to 6	12
6 up to 9	23
9 up to 12	8
12 up to 15	2
Total	50

a. How many employees accumulated less than 3,000 miles?
b. Convert the frequency distribution to a cumulative frequency distribution.
c. Portray the cumulative distribution in the form of a cumulative frequency polygon.
d. Based on the cumulative relative frequencies, about 75% of the employees accumulated how many miles or less?

22. The frequency distribution of order lead time of the retailer from Exercise 18 is repeated below.

Lead Time (days)	Frequency
0 up to 5	6
5 up to 10	7
10 up to 15	12
15 up to 20	8
20 up to 25	7
Total	40

a. How many orders were filled in less than 10 days? In less than 15 days?
b. Convert the frequency distribution to cumulative frequency and cumulative relative frequency distributions.
c. Develop a cumulative frequency polygon.
d. About 60% of the orders were filled in less than how many days?

CHAPTER SUMMARY

I. A frequency table is a grouping of qualitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.
II. A relative frequency table shows the fraction of the number of frequencies in each class.
III. A bar chart is a graphic representation of a frequency table.
IV. A pie chart shows the proportion each distinct class represents of the total number of observations.
V. A frequency distribution is a grouping of data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.
A. The steps in constructing a frequency distribution are
1. Decide on the number of classes.
2. Determine the class interval.
3. Set the individual class limits.
4. Tally the raw data into classes and determine the frequency in each class.

B. The class frequency is the number of observations in each class.
C. The class interval is the difference between the limits of two consecutive classes.
D. The class midpoint is halfway between the limits of consecutive classes.

VI. A relative frequency distribution shows the percent of observations in each class.

VII. There are several methods for graphically portraying a frequency distribution.

A. A histogram portrays the frequencies in the form of a rectangle or bar for each class. The height of the rectangles is proportional to the class frequencies.
B. A frequency polygon consists of line segments connecting the points formed by the intersection of the class midpoint and the class frequency.
C. A graph of a cumulative frequency distribution shows the number of observations less than a given value.
D. A graph of a cumulative relative frequency distribution shows the percent of observations less than a given value.

CHAPTER EXERCISES

23. Describe the similarities and differences of qualitative and quantitative variables. Be sure to include the following:
a. What level of measurement is required for each variable type?
b. Can both types be used to describe both samples and populations?

24. Describe the similarities and differences between a frequency table and a frequency distribution. Be sure to include which requires qualitative data and which requires quantitative data.

25. Alexandra Damonte will be building a new resort in Myrtle Beach, South Carolina. She must decide how to design the resort based on the type of activities that the resort will offer to its customers. A recent poll of 300 potential customers showed the following results about customers' preferences for planned resort activities:

Like planned activities	63
Do not like planned activities	135
Not sure	78
No answer	24

a. What is the table called?
b. Draw a bar chart to portray the survey results.
c. Draw a pie chart for the survey results.
d. If you are preparing to present the results to Ms. Damonte as part of a report, which graph would you prefer to show? Why?

26. FILE Speedy Swift is a package delivery service that serves the greater Atlanta, Georgia, metropolitan area. To maintain customer loyalty, one of Speedy Swift's performance objectives is on-time delivery. To monitor its performance, each delivery is measured on the following scale: early (package delivered before the promised time), on-time (package delivered within 5 minutes of the promised time), late (package delivered more than 5 minutes past the promised time), or lost (package never delivered). Speedy Swift's objective is to deliver 99% of all packages either early or on-time. Speedy collected the following data for last month's performance:

On-time	On-time	Early	Late	On-time	On-time	On-time	On-time	Late	On-time
Early	On-time	On-time	Early	On-time	On-time	On-time	On-time	On-time	On-time
Early	On-time	Early	On-time	On-time	On-time	Early	On-time	On-time	On-time
Early	On-time	On-time	Late	Early	Early	On-time	On-time	On-time	Early
On-time	Late	Late	On-time	On-time	On-time	On-time	On-time	On-time	On-time
On-time	Late	Early	On-time	Early	On-time	Lost	On-time	On-time	On-time
Early	Early	On-time	On-time	Late	Early	Lost	On-time	On-time	On-time
On-time	On-time	Early	On-time	Early	On-time	Early	On-time	Late	On-time
On-time	Early	On-time	On-time	On-time	Late	On-time	Early	On-time	On-time
On-time	On-time	On-time	On-time	On-time	Early	Early	On-time	On-time	On-time

a. What kind of variable is delivery performance? What scale is used to measure delivery performance?
b. Construct a frequency table for delivery performance for last month.
c. Construct a relative frequency table for delivery performance last month.
d. Construct a bar chart of the frequency table for delivery performance for last month.
e. Construct a pie chart of on-time delivery performance for last month.
f. Write a memo reporting the results of the analyses. Include your tables and graphs with written descriptions of what they show. Conclude with a general statement of last month's delivery performance as it relates to Speedy Swift's performance objectives.

27. A data set consists of 83 observations. How many classes would you recommend for a frequency distribution?

28. A data set consists of 145 observations that range from 56 to 490. What size class interval would you recommend?

29. FILE The following is the number of minutes to commute from home to work for a group of 25 automobile executives.

28	25	48	37	41	19	32	26	16	23	23	29	36
31	26	21	32	25	31	43	35	42	38	33	28	

a. How many classes would you recommend?
b. What class interval would you suggest?
c. What would you recommend as the lower limit of the first class?
d. Organize the data into a frequency distribution.
e. Comment on the shape of the frequency distribution.

30. FILE The following data give the weekly amounts spent on groceries for a sample of 45 households.

$271	$363	$159	$ 76	$227	$337	$295	$319	$250
279	205	279	266	199	177	162	232	303
192	181	321	309	246	278	50	41	335
116	100	151	240	474	297	170	188	320
429	294	570	342	279	235	434	123	325

a. How many classes would you recommend?
b. What class interval would you suggest?
c. What would you recommend as the lower limit of the first class?
d. Organize the data into a frequency distribution.

31. FILE A social scientist is studying the use of iPods by college students. A sample of 45 students revealed they played the following number of songs yesterday.

4	6	8	7	9	6	3	7	7	6	7	1	4	7	7
4	6	4	10	2	4	6	3	4	6	8	4	3	3	6
8	8	4	6	4	6	5	5	9	6	8	8	6	5	10

Organize the information into a frequency distribution.
a. How many classes would you suggest?
b. What is the most suitable class interval?
c. What is the lower limit of the initial class?
d. Create the frequency distribution.
e. Describe the shape of the distribution.

32. FILE David Wise handles his own investment portfolio, and has done so for many years. Listed below is the holding time (recorded to the nearest whole year) between purchase and sale for his collection of 36 stocks.

8	8	6	11	11	9	8	5	11	4	8	5	14	7	12	8	6	11	9	7
9	15	8	8	12	5	9	8	5	9	10	11	3	9	8	6				

a. How many classes would you propose?
b. What class interval would you suggest?
c. What quantity would you use for the lower limit of the initial class?

d. Using your responses to parts (a), (b), and (c), create a frequency distribution.
e. Describe the shape of the frequency distribution.

33. FILE You are exploring the music in your iTunes library. The total play counts over the past year for the 27 songs on your "smart playlist" are shown below. Make a frequency distribution of the counts and describe its shape. It is often claimed that a small fraction of a person's songs will account for most of their total plays. Does this seem to be the case here?

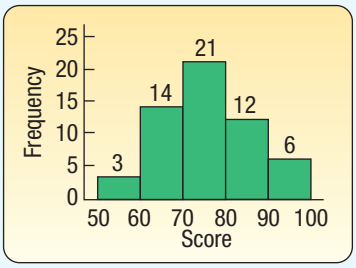
128	56	54	91	190	23	160	298	445	50
578	494	37	677	18	74	70	868	108	71
466	23	84	38	26	814	17			

34. FILE The monthly issues of the *Journal of Finance* are available on the Internet. The table below shows the number of times an issue was downloaded over the last 33 months. Suppose that you wish to summarize the number of downloads with a frequency distribution.

312	2,753	2,595	6,057	7,624	6,624	6,362	6,575	7,760	7,085	7,272
5,967	5,256	6,160	6,238	6,709	7,193	5,631	6,490	6,682	7,829	7,091
6,871	6,230	7,253	5,507	5,676	6,974	6,915	4,999	5,689	6,143	7,086

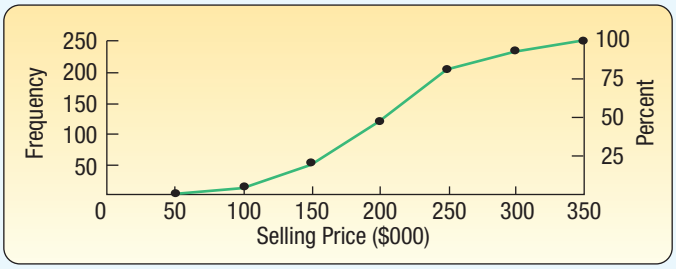
a. How many classes would you propose?
b. What class interval would you suggest?
c. What quantity would you use for the lower limit of the initial class?
d. Using your responses to parts (a), (b), and (c), create a frequency distribution.
e. Describe the shape of the frequency distribution.

35. The following histogram shows the scores on the first exam for a statistics class.



a. How many students took the exam?
b. What is the class interval?
c. What is the class midpoint for the first class?
d. How many students earned a score of less than 70?

36. The following chart summarizes the selling price of homes sold last month in the Sarasota, Florida, area.



a. What is the chart called?
b. How many homes were sold during the last month?
c. What is the class interval?
d. About 75% of the houses sold for less than what amount?
e. One hundred seventy-five of the homes sold for less than what amount?

37. FILE A chain of sport shops catering to beginning skiers, headquartered in Aspen, Colorado, plans to conduct a study of how much a beginning skier spends on his or her initial purchase of equipment and supplies. Based on these figures, it wants to explore the possibility of offering combinations, such as a pair of boots and a pair of skis, to induce customers to buy more. A sample of 44 cash register receipts revealed these initial purchases:

$140	$ 82	$265	$168	$ 90	$114	$172	$230	$142
86	125	235	212	171	149	156	162	118
139	149	132	105	162	126	216	195	127
161	135	172	220	229	129	87	128	126
175	127	149	126	121	118	172	126	

a. Arrive at a suggested class interval.
b. Organize the data into a frequency distribution using a lower limit of $70.
c. Interpret your findings.

38. FILE The numbers of outstanding shares for 24 publicly traded companies are listed in the following table.

Company	Number of Outstanding Shares (millions)	Company	Number of Outstanding Shares (millions)
Southwest Airlines	738	Costco	436
FirstEnergy	418	Home Depot	1,495
Harley Davidson	226	DTE Energy	172
Entergy	178	Dow Chemical	1,199
Chevron	1,957	Eastman Kodak	272
Pacific Gas and Electric	430	American Electric Power	485
DuPont	932	ITT Corporation	93
Westinghouse	22	Ameren	243
Eversource	314	Virginia Electric and Power	575
Facebook	1,067	Public Service Electric & Gas	506
Google, Inc.	64	Consumers Energy	265
Apple	941	Starbucks	744

a. Using the number of outstanding shares, summarize the companies with a frequency distribution.
b. Display the frequency distribution with a frequency polygon.
c. Create a cumulative frequency distribution of the outstanding shares.
d. Display the cumulative frequency distribution with a cumulative frequency polygon.
e. Based on the cumulative relative frequency distribution, 75% of the companies have less than "what number" of outstanding shares?
f. Write a brief analysis of this group of companies based on your statistical summaries of "number of outstanding shares."

39. A recent survey showed that the typical American car owner spends $2,950 per year on operating expenses. Below is a breakdown of the various expenditure items. Draw an appropriate chart to portray the data and summarize your findings in a brief report.

Expenditure Item	Amount
Fuel	$ 603
Interest on car loan	279
Repairs	930
Insurance and license	646
Depreciation	492
Total	$2,950

40. FILE Midland National Bank selected a sample of 40 student checking accounts. Below are their end-of-the-month balances.

$404	$ 74	$234	$149	$279	$215	$123	$ 55	$ 43	$321
87	234	68	489	57	185	141	758	72	863
703	125	350	440	37	252	27	521	302	127
968	712	503	489	327	608	358	425	303	203

a. Tally the data into a frequency distribution using $100 as a class interval and $0 as the starting point.
b. Draw a cumulative frequency polygon.
c. The bank considers any student with an ending balance of $400 or more a "preferred customer." Estimate the percentage of preferred customers.
d. The bank is also considering a service charge to the lowest 10% of the ending balances. What would you recommend as the cutoff point between those who have to pay a service charge and those who do not?

41. Residents of the state of South Carolina earned a total of $69.5 billion in adjusted gross income. Seventy-three percent of the total was in wages and salaries; 11% in dividends, interest, and capital gains; 8% in IRAs and taxable pensions; 3% in business income pensions; 2% in Social Security; and the remaining 3% from other sources. Develop a pie chart depicting the breakdown of adjusted gross income. Write a paragraph summarizing the information.

42. FILE A recent study of home technologies reported the number of hours of personal computer usage per week for a sample of 60 persons. Excluded from the study were people who worked out of their home and used the computer as a part of their work.

9.3	5.3	6.3	8.8	6.5	0.6	5.2	6.6	9.3	4.3
6.3	2.1	2.7	0.4	3.7	3.3	1.1	2.7	6.7	6.5
4.3	9.7	7.7	5.2	1.7	8.5	4.2	5.5	5.1	5.6
5.4	4.8	2.1	10.1	1.3	5.6	2.4	2.4	4.7	1.7
2.0	6.7	1.1	6.7	2.2	2.6	9.8	6.4	4.9	5.2
4.5	9.3	7.9	4.6	4.3	4.5	9.2	8.5	6.0	8.1

a. Organize the data into a frequency distribution. How many classes would you suggest? What value would you suggest for a class interval?
b. Draw a histogram. Describe your result.

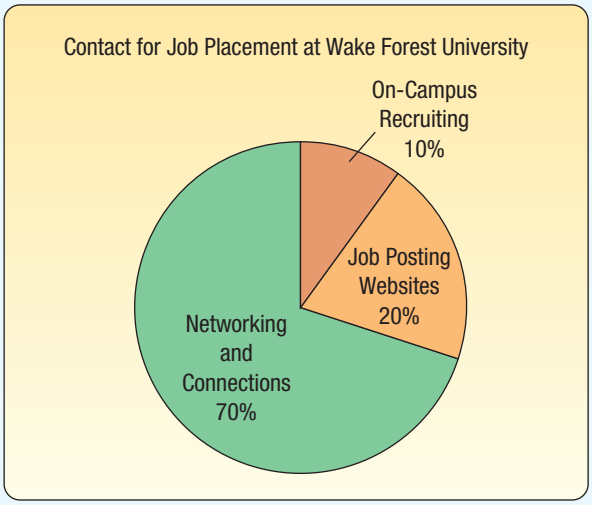
43. FILE Merrill Lynch recently completed a study regarding the size of online investment portfolios (stocks, bonds, mutual funds, and certificates of deposit) for a sample of clients in the 40 up to 50 years old age group. Listed following is the value of all the investments in thousands of dollars for the 70 participants in the study.

$669.9	$ 7.5	$ 77.2	$ 7.5	$125.7	$516.9	$ 219.9	$645.2
301.9	235.4	716.4	145.3	26.6	187.2	315.5	89.2
136.4	616.9	440.6	408.2	34.4	296.1	185.4	526.3
380.7	3.3	363.2	51.9	52.2	107.5	82.9	63.0
228.6	308.7	126.7	430.3	82.0	227.0	321.1	403.4
39.5	124.3	118.1	23.9	352.8	156.7	276.3	23.5
31.3	301.2	35.7	154.9	174.3	100.6	236.7	171.9
221.1	43.4	212.3	243.3	315.4	5.9	1,002.2	171.7
295.7	437.0	87.8	302.1	268.1	899.5		

a. Organize the data into a frequency distribution. How many classes would you suggest? What value would you suggest for a class interval?
b. Draw a histogram. Financial experts suggest that this age group of people have at least five times their salary saved. As a benchmark, assume an investment portfolio of $500,000 would support retirement in 10–15 years. In writing, summarize your results.

44. A total of 5.9% of the prime-time viewing audience watched shows on ABC, 7.6% watched shows on CBS, 5.5% on Fox, 6.0% on NBC, 2.0% on Warner Brothers, and 2.2% on UPN. A total of 70.8% of the audience watched shows on other cable networks, such as CNN and ESPN. You can find the latest information on TV viewing from the following website: **http://www.nielsen.com/us/en/top10s.html/**. Develop a pie chart or a bar chart to depict this information. Write a paragraph summarizing your findings.

45. Refer to the following chart:

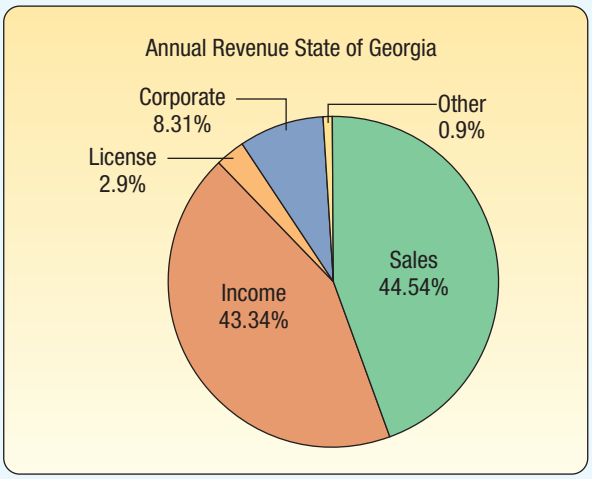

Contact for Job Placement at Wake Forest University
On-Campus Recruiting 10%
Job Posting Websites 20%
Networking and Connections 70%

a. What is the name given to this type of chart?

b. Suppose that 1,000 graduates will start a new job shortly after graduation. Estimate the number of graduates whose first contact for employment occurred through networking and other connections.

c. Would it be reasonable to conclude that about 90% of job placements were made through networking, connections, and job posting websites? Cite evidence.

46. The following chart depicts the annual revenues, by type of tax, for the state of Georgia.



a. What percentage of the state revenue is accounted for by sales tax and individual income tax?

b. Which category will generate more revenue: corporate taxes or license fees?

c. The total annual revenue for the state of Georgia is $6.3 billion. Estimate the amount of revenue in billions of dollars for sales taxes and for individual taxes.

47. In 2014, the United States exported a total of $376 billion worth of products to Canada. The five largest categories were:

Product	Amount
Vehicles	$63.3
Machinery	59.7
Electrical machinery	36.6
Mineral fuel and oil	24.8
Plastic	17.0

a. Use a software package to develop a bar chart.
b. What percentage of the United States' total exports to Canada is represented by the two categories "Machinery" and "Electrical Machinery"?
c. What percentage of the top five exported products do "Machinery" and "Electrical Machinery" represent?

48. FILE In the United States, the industrial revolution of the early 20th century changed farming by making it more efficient. For example, in 1910 U.S. farms used 24.2 million horses and mules and only about 1,000 tractors. By 1960, 4.6 million tractors were used and only 3.2 million horses and mules. An outcome of making farming more efficient is the reduction of the number of farms from over 6 million in 1920 to about 2.2 million farms today. Listed below is the number of farms, in thousands, for each of the 50 states. Summarize the data and write a paragraph that describes your findings.

50	12	5	28	59	19	35	22	80	5
8	48	3	75	25	77	46	68	10	69
77	25	13	20	35	6	52	61	36	38
88	1	75	246	59	50	44	98	74	2
32	42	7	31	28	9	8	44	25	37

49. One of the most popular candies in the United States is M&M's produced by the Mars Company. In the beginning M&M's were all brown. Now they are produced in red, green, blue, orange, brown, and yellow. Recently, the purchase of a 14-ounce bag of M&M's Plain had 444 candies with the following breakdown by color: 130 brown, 98 yellow, 96 red, 35 orange, 52 blue, and 33 green. Develop a chart depicting this information and write a paragraph summarizing the results.

50. FILE The number of families who used the Minneapolis YWCA day care service was recorded during a 30-day period. The results are as follows:

31	49	19	62	24	45	23	51	55	60
40	35	54	26	57	37	43	65	18	41
50	56	4	54	39	52	35	51	63	42

a. Construct a cumulative frequency distribution.
b. Sketch a graph of the cumulative frequency polygon.
c. How many days saw fewer than 30 families utilize the day care center?
d. Based on cumulative relative frequencies, how busy were the highest 80% of the days?

DATA ANALYTICS

51. FILE Refer to the North Valley Real Estate data that reports information on homes sold during the last year. For the variable *price*, select an appropriate class interval and organize the selling prices into a frequency distribution. Write a brief report summarizing your findings. Be sure to answer the following questions in your report.
a. Around what values of price do the data tend to cluster?
b. Based on the frequency distribution, what is the typical selling price in the first class? What is the typical selling price in the last class?

c. Draw a cumulative relative frequency distribution. Using this distribution, fifty percent of the homes sold for what price or less? Estimate the lower price of the top ten percent of homes sold. About what percent of the homes sold for less than $300,000?

d. Refer to the variable *bedrooms*. Draw a bar chart showing the number of homes sold with 2, 3, 4 or more bedrooms. Write a description of the distribution.

52. FILE Refer to the Baseball 2016 data that report information on the 30 Major League Baseball teams for the 2016 season. Create a frequency distribution for the *Team Salary* variable and answer the following questions.

a. What is the typical salary for a team? What is the range of the salaries?

b. Comment on the shape of the distribution. Does it appear that any of the teams have a salary that is out of line with the others?

c. Draw a cumulative relative frequency distribution of team salary. Using this distribution, forty percent of the teams have a salary of less than what amount? About how many teams have a total salary of more than $220 million?

53. FILE Refer to the Lincolnville School District bus data. Select the variable referring to the number of *miles traveled since the last maintenance*, and then organize these data into a frequency distribution.

a. What is a typical amount of miles traveled? What is the range?

b. Comment on the shape of the distribution. Are there any outliers in terms of miles driven?

c. Draw a cumulative relative frequency distribution. Forty percent of the buses were driven fewer than how many miles? How many buses were driven less than 10,500 miles?

d. Refer to the variables regarding the bus *manufacturer* and the bus *capacity.* Draw a pie chart of each variable and write a description of your results.

Describing Data:

NUMERICAL MEASURES

3

© Andy Lyons/Getty Images

▲ **THE KENTUCKY DERBY** is held the first Saturday in May at Churchill Downs in Louisville, Kentucky. The race track is one and one-quarter miles. The table in Exercise 82 shows the winners since 1990, their margin of victory, the winning time, and the payoff on a $2 bet. Determine the mean and median for the variables winning time and payoff on a $2 bet. (See Exercise 82 and LO3-1.)

LEARNING OBJECTIVES

When you have completed this chapter, you will be able to:

LO3-1 Compute and interpret the mean, the median, and the mode.

LO3-2 Compute a weighted mean.

LO3-3 Compute and interpret the geometric mean.

LO3-4 Compute and interpret the range, variance, and standard deviation.

LO3-5 Explain and apply Chebyshev's theorem and the Empirical Rule.

LO3-6 Compute the mean and standard deviation of grouped data.

INTRODUCTION

STATISTICS IN ACTION

Did you ever meet the "average" American man? Well, his name is Robert (that is the nominal level of measurement) and he is 31 years old (that is the ratio level), is 69.5 inches tall (again the ratio level of measurement), weighs 172 pounds, wears a size 9½ shoe, has a 34-inch waist, and wears a size 40 suit. In addition, the average man eats 4 pounds of potato chips, watches 1,456 hours of TV, and eats 26 pounds of bananas each year, and also sleeps 7.7 hours per night.

The average American woman is 5′ 4″ tall and weighs 140 pounds, while the average American model is 5′ 11″ tall and weighs 117 pounds. On any given day, almost half of the women in the United States are on a diet. Idolized in the 1950s, Marilyn Monroe would be considered overweight by today's standards. She fluctuated between a size 14 and a size 18 dress, and was a healthy and attractive woman.

Chapter 2 began our study of descriptive statistics. To summarize raw data into a meaningful form, we organized qualitative data into a frequency table and portrayed the results in a bar chart. In a similar fashion, we organized quantitative data into a frequency distribution and portrayed the results in a histogram. We also looked at other graphical techniques such as pie charts to portray qualitative data and frequency polygons to portray quantitative data.

This chapter is concerned with two numerical ways of describing quantitative variables, namely, **measures of location** and **measures of dispersion.** Measures of location are often referred to as averages. The purpose of a measure of location is to pinpoint the center of a distribution of data. An average is a measure of location that shows the central value of the data. Averages appear daily on TV, on various websites, in the newspaper, and in other journals. Here are some examples:

- The average U.S. home changes ownership every 11.8 years.
- An American receives an average of 568 pieces of mail per year.
- The average American home has more TV sets than people. There are 2.73 TV sets and 2.55 people in the typical home.
- The average American couple spends $20,398 for their wedding, while their budget is 50% less. This does not include the cost of a honeymoon or engagement ring.
- The average price of a theater ticket in the United States is $8.31, according to the National Association of Theater Owners.

© Andersen Ross/Getty Images RF

If we consider only measures of location in a set of data, or if we compare several sets of data using central values, we may draw an erroneous conclusion. In addition to measures of location, we should consider the **dispersion**—often called the *variation* or the *spread*—in the data. As an illustration, suppose the average annual income of executives for Internet-related companies is $80,000, and the average income for executives in pharmaceutical firms is also $80,000. If we looked only at the average incomes, we might wrongly conclude that the distributions of the two salaries are the same. However, we need to examine the dispersion or spread of the distributions of salary. A look at the salary ranges indicates that this conclusion of equal distributions is not correct. The salaries for the executives in the Internet firms range from $70,000 to $90,000, but salaries for the marketing executives in pharmaceuticals range from $40,000 to $120,000. Thus, we conclude that although the average salaries are the same for the two industries, there is much more spread or dispersion in salaries for the pharmaceutical executives. To describe the dispersion, we will consider the range, the variance, and the standard deviation.

LO3-1
Compute and interpret the mean, the median, and the mode.

MEASURES OF LOCATION

We begin by discussing measures of location. There is not just one measure of location; in fact, there are many. We will consider five: the arithmetic mean, the median, the mode, the weighted mean, and the geometric mean. The arithmetic mean is the most widely used and widely reported measure of location. We study the mean as both a population parameter and a sample statistic.

The Population Mean

Many studies involve all the values in a population. For example, there are 12 sales associates employed at the Reynolds Road Carpet Outlet. The mean amount of commission they earned last month was $1,345. This is a population value because we considered the commission of *all* the sales associates. Other examples of a population mean would be:

- The mean closing price for Johnson & Johnson stock for the last 5 days is $95.47.
- The mean number of hours of overtime worked last week by the six welders in the welding department of Butts Welding Inc. is 6.45 hours.
- Caryn Tirsch began a website last month devoted to organic gardening. The mean number of hits on her site for the 31 days in July was 84.36.

For raw data—that is, data that have not been grouped in a frequency distribution—the population mean is the sum of all the values in the population divided by the number of values in the population. To find the population mean, we use the following formula.

$$\text{Population mean} = \frac{\text{Sum of all the values in the population}}{\text{Number of values in the population}}$$

Instead of writing out in words the full directions for computing the population mean (or any other measure), it is more convenient to use the shorthand symbols of mathematics. The mean of the population using mathematical symbols is:

POPULATION MEAN

$$\mu = \frac{\Sigma x}{N} \qquad (3\text{–}1)$$

where:

- μ represents the population mean. It is the Greek lowercase letter "mu."
- N is the number of values in the population.
- x represents any particular value.
- Σ is the Greek capital letter "sigma" and indicates the operation of adding.
- Σx is the sum of the x values in the population.

Any measurable characteristic of a population is called a **parameter.** The mean of a population is an example of a parameter.

PARAMETER A characteristic of a population.

EXAMPLE

There are 42 exits on I-75 through the state of Kentucky. Listed below are the distances between exits (in miles).

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

Why is this information a population? What is the mean number of miles between exits?

SOLUTION

This is a population because we are considering all the exits on I-75 in Kentucky. We add the distances between each of the 42 exits. The total distance is 192 miles. To find the arithmetic mean, we divide this total by 42. So the arithmetic mean is 4.57 miles, found by 192/42. From formula (3–1):

$$\mu = \frac{\Sigma x}{N} = \frac{11 + 4 + 10 + \cdots + 1}{42} = \frac{192}{42} = 4.57$$

How do we interpret the value of 4.57? It is the typical number of miles between exits. Because we considered all the exits on I-75 in Kentucky, this value is a population parameter.

© Bloomberg/Getty Images

The Sample Mean

As explained in Chapter 1, we often select a sample from the population to estimate a specific characteristic of the population. Smucker's quality assurance department needs to be assured that the amount of orange marmalade in the jar labeled as containing 12 ounces actually contains that amount. It would be very expensive and time-consuming to check the weight of each jar. Therefore, a sample of 20 jars is selected, the mean of the sample is determined, and that value is used to estimate the amount in each jar.

For raw data—that is, ungrouped data—*the mean is the sum of all the sampled values divided by the total number of sampled values.* To find the mean for a sample:

$$\text{Sample mean} = \frac{\text{Sum of all the values in the sample}}{\text{Number of values in the sample}}$$

The mean of a sample and the mean of a population are computed in the same way, but the shorthand notation used is different. The formula for the mean of a *sample* is:

SAMPLE MEAN $$\bar{x} = \frac{\Sigma x}{n} \qquad \textbf{(3–2)}$$

where:

- $\bar{x}$ represents the sample mean. It is read "x bar."
- n is the number of values in the sample.
- x represents any particular value.
- Σ is the Greek capital letter "sigma" and indicates the operation of adding.
- Σx is the sum of the x values in the sample.

The mean of a sample, or any other measure based on sample data, is called a **statistic.** If the mean weight of a sample of 10 jars of Smucker's orange marmalade is 11.5 ounces, this is an example of a statistic.

STATISTIC A characteristic of a sample.

EXAMPLE

Verizon is studying the number of monthly minutes used by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month.

90	77	94	89	119	112
91	110	92	100	113	83

What is the arithmetic mean number of minutes used last month?

SOLUTION

Using formula (3–2), the sample mean is:

$$\text{Sample mean} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{90 + 77 + \cdots + 83}{12} = \frac{1{,}170}{12} = 97.5$$

The arithmetic mean number of minutes used last month by the sample of cell phone users is 97.5 minutes.

Properties of the Arithmetic Mean

The arithmetic mean is a widely used measure of location. It has several important properties:

1. **To compute a mean, the data must be measured at the interval or ratio level.** Recall from Chapter 1 that ratio-level data include such data as ages, incomes, and weights, with the distance between numbers being constant.
2. **All the values are included in computing the mean.**
3. **The mean is unique.** That is, there is only one mean in a set of data. Later in the chapter, we will discover a measure of location that might appear twice, or more than twice, in a set of data.
4. **The sum of the deviations of each value from the mean is zero.** Expressed symbolically:

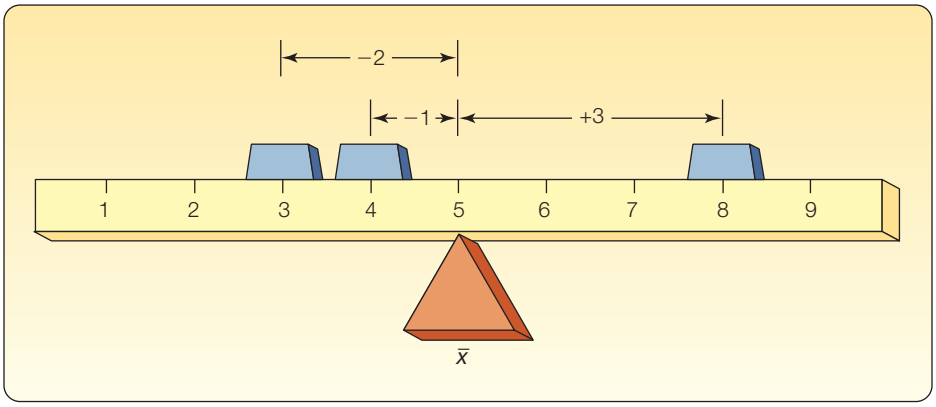
$$\Sigma(x - \bar{x}) = 0$$

As an example, the mean of 3, 8, and 4 is 5. Then:

$$\begin{aligned}\Sigma(x - \bar{x}) &= (3 - 5) + (8 - 5) + (4 - 5)\\ &= -2 + 3 - 1\\ &= 0\end{aligned}$$

Thus, we can consider the mean as a balance point for a set of data. To illustrate, we have a long board with the numbers 1, 2, 3, . . . , 9 evenly spaced on it. Suppose three bars of equal weight were placed on the board at numbers 3, 4, and 8, and the balance point was set at 5, the mean of the three numbers. We would find that the

board is balanced perfectly! The deviations below the mean (−3) are equal to the deviations above the mean (+3). Shown schematically:



The mean does have a weakness. Recall that the mean uses the value of every item in a sample, or population, in its computation. If one or two of these values are either extremely large or extremely small compared to the majority of data, the mean might not be an appropriate average to represent the data. For example, suppose the annual incomes of a sample of financial planners at Merrill Lynch are $62,900, $61,600, $62,500, $60,800, and $1,200,000. The mean income is $289,560. Obviously, it is not representative of this group because all but one financial planner has an income in the $60,000 to $63,000 range. One income ($1.2 million) is unduly affecting the mean.

SELF-REVIEW 3–1

1. The annual incomes of a sample of middle-management employees at Westinghouse are $62,900, $69,100, $58,300, and $76,800.
   (a) Give the formula for the sample mean.
   (b) Find the sample mean.
   (c) Is the mean you computed in (b) a statistic or a parameter? Why?
   (d) What is your best estimate of the population mean?
2. The six students in Computer Science 411 are a population. Their final course grades are 92, 96, 61, 86, 79, and 84.
   (a) Give the formula for the population mean.
   (b) Compute the mean course grade.
   (c) Is the mean you computed in (b) a statistic or a parameter? Why?

EXERCISES

The answers to the odd-numbered exercises are in Appendix D.

1. Compute the mean of the following population values: 6, 3, 5, 7, 6.
2. Compute the mean of the following population values: 7, 5, 7, 3, 7, 4.
3. **a.** Compute the mean of the following sample values: 5, 9, 4, 10.
b. Show that $\Sigma(x - \bar{x}) = 0$.
4. **a.** Compute the mean of the following sample values: 1.3, 7.0, 3.6, 4.1, 5.0.
b. Show that $\Sigma(x - \bar{x}) = 0$.
5. Compute the mean of the following sample values: 16.25, 12.91, 14.58.
6. Suppose you go to the grocery store and spend $61.85 for the purchase of 14 items. What is the mean price per item?

For Exercises 7–10, (a) compute the arithmetic mean and (b) indicate whether it is a statistic or a parameter.

7. There are 10 salespeople employed by Midtown Ford. The number of new cars sold last month by the respective salespeople were: 15, 23, 4, 19, 18, 10, 10, 8, 28, 19.
8. A mail-order company counted the number of incoming calls per day to the company's toll-free number during the first 7 days in May: 14, 24, 19, 31, 36, 26, 17.
9. FILE The Cambridge Power and Light Company selected a random sample of 20 residential customers. Following are the amounts, to the nearest dollar, the customers were charged for electrical service last month:

54	48	58	50	25	47	75	46	60	70
67	68	39	35	56	66	33	62	65	67

10. FILE A Human Resources manager at Metal Technologies studied the overtime hours of welders. A sample of 15 welders showed the following number of overtime hours worked last month.

13	13	12	15	7	15	5	12
6	7	12	10	9	13	12	

11. AAA Heating and Air Conditioning completed 30 jobs last month with a mean revenue of $5,430 per job. The president wants to know the total revenue for the month. Based on the limited information, can you compute the total revenue? What is it?
12. A large pharmaceutical company hires business administration graduates to sell its products. The company is growing rapidly and dedicates only 1 day of sales training for new salespeople. The company's goal for new salespeople is $10,000 per month. The goal is based on the current mean sales for the entire company, which is $10,000 per month. After reviewing the retention rates of new employees, the company finds that only 1 in 10 new employees stays longer than 3 months. Comment on using the current mean sales per month as a sales goal for new employees. Why do new employees leave the company?

The Median

We have stressed that, for data containing one or two very large or very small values, the arithmetic mean may not be representative. The center for such data is better described by a measure of location called the **median.**

To illustrate the need for a measure of location other than the arithmetic mean, suppose you are seeking to buy a condominium in Palm Aire. Your real estate agent says that the typical price of the units currently available is $110,000. Would you still want to look? If you had budgeted your maximum purchase price at $75,000, you might think they are out of your price range. However, checking the prices of the individual units might change your mind. They are $60,000, $65,000, $70,000, and $80,000, and a superdeluxe penthouse costs $275,000. The arithmetic mean price is $110,000, as the real estate agent reported, but one price ($275,000) is pulling the arithmetic mean upward, causing it to be an unrepresentative average. It does seem that a price around $70,000 is a more typical or representative average, and it is. In cases such as this, the median provides a more valid measure of location.

MEDIAN The midpoint of the values after they have been ordered from the minimum to the maximum values.

The median price of the units available is $70,000. To determine this, we order the prices from the minimum value ($60,000) to the maximum value ($275,000) and select the middle value ($70,000). For the median, the data must be at least an ordinal level of measurement.

Prices Ordered from Minimum to Maximum		Prices Ordered from Maximum to Minimum
$ 60,000		$275,000
65,000		80,000
70,000	← Median →	70,000
80,000		65,000
275,000		60,000

Note that there is the same number of prices below the median of $70,000 as above it. The median is, therefore, unaffected by extremely low or high prices. Had the highest price been $90,000, or $300,000, or even $1 million, the median price would still be $70,000. Likewise, had the lowest price been $20,000 or $50,000, the median price would still be $70,000.

In the previous illustration, there are an *odd* number of observations (five). How is the median determined for an *even* number of observations? As before, the observations are ordered. Then by convention to obtain a unique value we calculate the mean of the two middle observations. So for an even number of observations, the median may not be one of the given values.

EXAMPLE

Facebook is a popular social networking website. Users can add friends and send them messages, and update their personal profiles to notify friends about themselves and their activities. A sample of 10 adults revealed they spent the following number of hours last month using Facebook.

3	5	7	5	9	1	3	9	17	10

Find the median number of hours.

SOLUTION

Note that the number of adults sampled is even (10). The first step, as before, is to order the hours using Facebook from the minimum value to the maximum value. Then identify the two middle times. The arithmetic mean of the two middle observations gives us the median hours. Arranging the values from minimum to maximum:

1	3	3	5	5	7	9	9	10	17

The median is found by averaging the two middle values. The middle values are 5 hours and 7 hours, and the mean of these two values is 6. We conclude that the typical adult Facebook user spends 6 hours per month at the website. Notice that the median is not one of the values. Also, half of the times are below the median and half are above it.

The major properties of the median are:

1. **It is not affected by extremely large or small values.** Therefore, the median is a valuable measure of location when such values do occur.
2. **It can be computed for ordinal-level data or higher.** Recall from Chapter 1 that ordinal-level data can be ranked from low to high.

The Mode

The **mode** is another measure of location.

MODE The value of the observation that appears most frequently.

The mode is especially useful in summarizing nominal-level data. As an example of its use for nominal-level data, a company has developed five bath oils. The bar chart in Chart 3–1 shows the results of a marketing survey designed to find which bath oil consumers prefer. The largest number of respondents favored Lamoure, as evidenced by the highest bar. Thus, Lamoure is the mode.

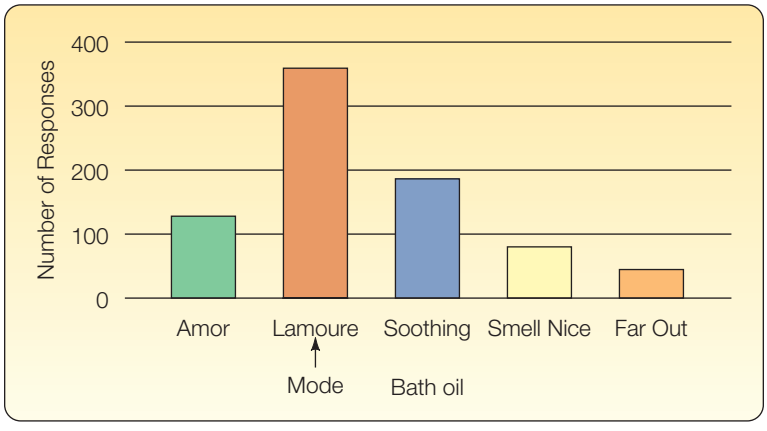


CHART 3–1 Number of Respondents Favoring Various Bath Oils

EXAMPLE

Recall the data regarding the distance in miles between exits on I-75 in Kentucky. The information is repeated below.

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

What is the modal distance?

SOLUTION

The first step is to organize the distances into a frequency table. This will help us determine the distance that occurs most frequently.

Distance in Miles between Exits	Frequency
1	8
2	7
3	7
4	3
5	4
6	1
7	3
8	2
9	1
10	4
11	1
14	1
Total	42

The distance that occurs most often is 1 mile. This happens eight times—that is, there are eight exits that are 1 mile apart. So the modal distance between exits is 1 mile.

Which of the three measures of location (mean, median, or mode) best represents the central location of these data? Is the mode the best measure of location to represent the Kentucky data? No. The mode assumes only the nominal scale of measurement and the variable miles is measured using the ratio scale. We calculated the mean to be 4.57 miles. See page 54. Is the mean the best measure of location to represent these data? Probably not. There are several cases in which the distance between exits is large. These values are affecting the mean, making it too large and not representative of the distances between exits. What about the median? The median distance is 3 miles. That is, half of the distances between exits are 3 miles or less. In this case, the median of 3 miles between exits is probably a more representative measure of the distance between exits.

In summary, we can determine the mode for all levels of data—nominal, ordinal, interval, and ratio. The mode also has the advantage of not being affected by extremely high or low values.

The mode does have disadvantages, however, that cause it to be used less frequently than the mean or median. For many sets of data, there is no mode because no value appears more than once. For example, there is no mode for this set of price data because every value occurs once: $19, $21, $23, $20, and $18. Conversely, for some data sets there is more than one mode. Suppose the ages of the individuals in a stock investment club are 22, 26, 27, 27, 31, 35, and 35. Both the ages 27 and 35 are modes. Thus, this grouping of ages is referred to as *bimodal* (having two modes). One would question the use of two modes to represent the location of this set of age data.

SELF-REVIEW 3–2

1. A sample of single persons in Towson, Texas, receiving Social Security payments revealed these monthly benefits: $852, $598, $580, $1,374, $960, $878, and $1,130.
   (a) What is the median monthly benefit?
   (b) How many observations are below the median? Above it?
2. The number of work stoppages in the United States over the last 10 years are 22, 20, 21, 15, 5, 11, 19, 19, 15, and 11.
   (a) What is the median number of stoppages?
   (b) How many observations are below the median? Above it?
   (c) What is the modal number of work stoppages?

EXERCISES

13. What would you report as the modal value for a set of observations if there were a total of:
- **a.** 10 observations and no two values were the same?
- **b.** 6 observations and they were all the same?
- **c.** 6 observations and the values were 1, 2, 3, 3, 4, and 4?

For Exercises 14–16, determine the (a) mean, (b) median, and (c) mode.

14. The following is the number of oil changes for the last 7 days at the Jiffy Lube located at the corner of Elm Street and Pennsylvania Avenue.

41 15 39 54 31 15 33

15. The following is the percent change in net income from last year to this year for a sample of 12 construction companies in Denver.

5 1 −10 −6 5 12 7 8 6 5 −1 11

16. The following are the ages of the 10 people in the Java Coffee Shop at the Southwyck Shopping Mall at 10 a.m.

21 41 20 23 24 33 37 42 23 29

17. FILE Several indicators of long-term economic growth in the United States and their annual percent change are listed below.

Economic Indicator	Percent Change	Economic Indicator	Percent Change
Inflation	4.5%	Real GNP	2.9%
Exports	4.7	Investment (residential)	3.6
Imports	2.3	Investment (nonresidential)	2.1
Real disposable income	2.9	Productivity (total)	1.4
Consumption	2.7	Productivity (manufacturing)	5.2

- **a.** What is the median percent change?
- **b.** What is the modal percent change?

18. FILE Sally Reynolds sells real estate along the coastal area of Northern California. Below are her total annual commissions between 2005 and 2015. Find the mean, median, and mode of the commissions she earned for the 11 years.

Year	Amount (thousands)
2005	292.16
2006	233.80
2007	206.97
2008	202.67
2009	164.69
2010	206.53
2011	237.51
2012	225.57
2013	255.33
2014	248.14
2015	269.11

19. FILE The accounting firm of Rowatti and Koppel specializes in income tax returns for self-employed professionals, such as physicians, dentists, architects, and lawyers. The firm employs 11 accountants who prepare the returns. For last year, the number of returns prepared by each accountant was:

58 75 31 58 46 65 60 71 45 58 80

Find the mean, median, and mode for the number of returns prepared by each accountant. If you could report only one, which measure of location would you recommend reporting?

20. FILE The demand for the video games provided by Mid-Tech Video Games Inc. has exploded in the last several years. Hence, the owner needs to hire several new technical people to keep up with the demand. Mid-Tech gives each applicant a special test that Dr. McGraw, the designer of the test, believes is closely related to the ability to create video games. For the general population, the mean on this test is 100. Below are the scores on this test for the applicants.

95	105	120	81	90	115	99	100	130	10

The president is interested in the overall quality of the job applicants based on this test. Compute the mean and the median scores for the 10 applicants. What would you report to the president? Does it seem that the applicants are better than the general population?

The Relative Positions of the Mean, Median, and Mode

Refer to the histogram in Chart 3–2. It is a symmetric distribution, which is also mound-shaped. This distribution *has the same shape on either side of the center*. If the histogram were folded in half, the two halves would be identical. For any symmetric distribution, the mode, median, and mean are located at the center and are always equal. They are all equal to 30 years in Chart 3–2. We should point out that there are symmetric distributions that are not mound-shaped.

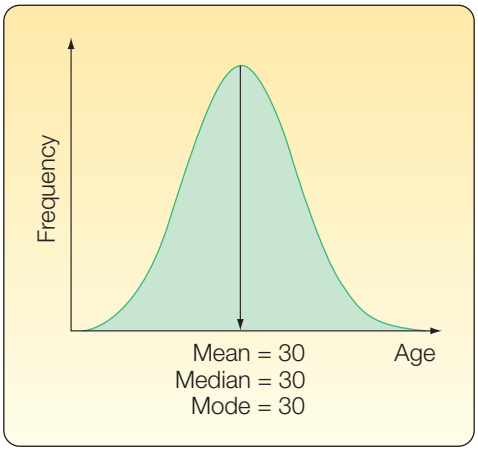


CHART 3–2 A Symmetric Distribution

The number of years corresponding to the highest point of the curve is the *mode* (30 years). Because the distribution is symmetrical, the *median* corresponds to the point where the distribution is cut in half (30 years). Also, because the arithmetic mean is the balance point of a distribution (as shown in the Properties of the Arithmetic Mean section on page 56), and the distribution is symmetric, the arithmetic mean is 30. Logically, any of the three measures would be appropriate to represent the distribution's center.

If a distribution is nonsymmetrical, or **skewed,** the relationship among the three measures changes. In a **positively skewed distribution,** such as the distribution of weekly income in Chart 3–3, the arithmetic mean is the largest of the three measures. Why? Because the mean is influenced more than the median or mode by a few extremely high values. The median is generally the next largest measure in a positively skewed frequency distribution. The mode is the smallest of the three measures.

If the distribution is highly skewed, the mean would not be a good measure to use. The median and mode would be more representative.

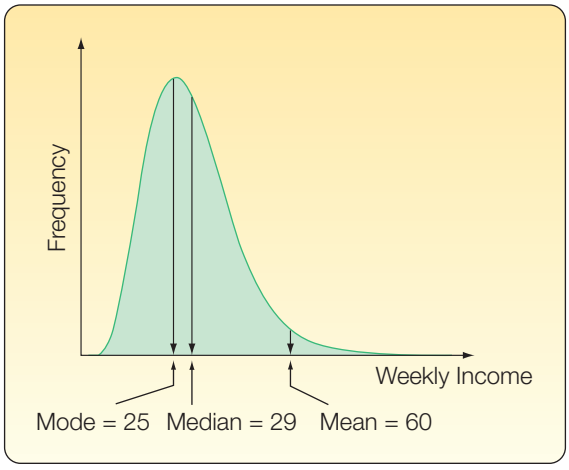


CHART 3–3 A Positively Skewed Distribution

Conversely, if a distribution is **negatively skewed,** such as the distribution of tensile strength in Chart 3–4, the mean is the lowest of the three measures. The mean is, of course, influenced by a few extremely low observations. The median is greater than the arithmetic mean, and the modal value is the largest of the three measures. Again, if the distribution is highly skewed, the mean should not be used to represent the data.

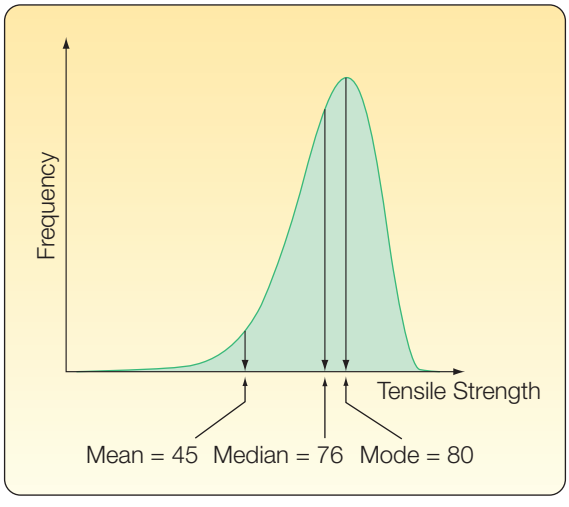


CHART 3–4 A Negatively Skewed Distribution

SELF-REVIEW 3–3

The weekly sales from a sample of Hi-Tec electronic supply stores were organized into a frequency distribution. The mean of weekly sales was computed to be $105,900, the median $105,000, and the mode $104,500.

(a) Sketch the sales in the form of a smoothed frequency polygon. Note the location of the mean, median, and mode on the *X*-axis.

(b) Is the distribution symmetrical, positively skewed, or negatively skewed? Explain.

EXERCISES

21. FILE The unemployment rate in the state of Alaska by month is given in the table below:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
8.7	8.8	8.7	7.8	7.3	7.8	6.6	6.5	6.5	6.8	7.3	7.6

a. What is the arithmetic mean of the Alaska unemployment rates?
b. Find the median and the mode for the unemployment rates.
c. Compute the arithmetic mean and median for just the winter (Dec–Mar) months. Is it much different?

22. FILE Big Orange Trucking is designing an information system for use in "in-cab" communications. It must summarize data from eight sites throughout a region to describe typical conditions. Compute an appropriate measure of central location for the variables wind direction, temperature, and pavement.

City	Wind Direction	Temperature	Pavement
Anniston, AL	West	89	Dry
Atlanta, GA	Northwest	86	Wet
Augusta, GA	Southwest	92	Wet
Birmingham, AL	South	91	Dry
Jackson, MS	Southwest	92	Dry
Meridian, MS	South	92	Trace
Monroe, LA	Southwest	93	Wet
Tuscaloosa, AL	Southwest	93	Trace

Software Solution

We can use a statistical software package to find many measures of location.

EXAMPLE

Table 2–4 on page 26 shows the profit on the sales of 180 vehicles at Applewood Auto Group. Determine the mean and the median selling price.

SOLUTION

The mean, median, and modal amounts of profit are reported in the following output (highlighted in the screen shot). (Reminder: The instructions to create the output appear in the **Software Commands** in Appendix C.) There are 180 vehicles in the study, so using a calculator would be tedious and prone to error.

APPLEWOOD AUTO GROUP

	A	B	C	D	E	F	G	H
1	Age	Profit	Location	Vehicle-Type	Previous		*Profit*	
2	21	$1,387	Tionesta	Sedan	0			
3	23	$1,754	Sheffield	SUV	1		Mean	1843.17
4	24	$1,817	Sheffield	Hybrid	1		Standard Error	47.97
5	25	$1,040	Sheffield	Compact	0		Median	1882.50
6	26	$1,273	Kane	Sedan	1		Mode	1915.00
7	27	$1,529	Sheffield	Sedan	1		Standard Deviation	643.63
8	27	$3,082	Kane	Truck	0		Sample Variance	414256.61
9	28	$1,951	Kane	SUV	1		Kurtosis	-0.22
10	28	$2,692	Tionesta	Compact	0		Skewness	-0.24
11	29	$1,342	Kane	Sedan	2		Range	2998
12	29	$1,206	Sheffield	Sedan	0		Minimum	294
13	30	$443	Kane	Sedan	3		Maximum	3292
14	30	$1,621	Sheffield	Truck	1		Sum	331770
15	30	$754	Olean	Sedan	2		Count	180

The mean profit is $1,843.17 and the median is $1,882.50. These two values are less than $40 apart, so either value is reasonable. We can also see from the Excel output that there were 180 vehicles sold and their total profit was $331,770.00. We will describe the meaning of standard error, standard deviation, and other measures reported on the output later in this chapter and in later chapters.

What can we conclude? The typical profit on a vehicle is about $1,850. Management at Applewood might use this value for revenue projections. For example, if the dealership could increase the number of vehicles sold in a month from 180 to 200, this would result in an additional estimated $37,000 of revenue, found by 20($1,850).

LO3-2
Compute a weighted mean.

THE WEIGHTED MEAN

The weighted mean is a convenient way to compute the arithmetic mean when there are several observations of the same value. To explain, suppose the nearby Wendy's Restaurant sold medium, large, and Biggie-sized soft drinks for $1.84, $2.07, and $2.40, respectively. Of the last 10 drinks sold, 3 were medium, 4 were large, and 3 were Biggie-sized. To find the mean price of the last 10 drinks sold, we could use formula (3–2).

$$\bar{x} = \frac{\$1.84 + \$1.84 + \$1.84 + \$2.07 + \$2.07 + \$2.07 + \$2.07 + \$2.40 + \$2.40 + \$2.40}{10}$$

$$\bar{x} = \frac{\$21.00}{10} = \$2.10$$

The mean selling price of the last 10 drinks is $2.10.

An easier way to find the mean selling price is to determine the weighted mean. That is, we multiply each observation by the number of times it occurs. We will refer to the weighted mean as $\bar{x}_w$. This is read "x bar sub w."

$$\bar{x}_w = \frac{3(\$1.84) + 4(\$2.07) + 3(\$2.40)}{10} = \frac{\$21.00}{10} = \$2.10$$

In this case, the weights are frequency counts. However, any measure of importance could be used as a weight. In general, the weighted mean of a set of numbers designated $x_1, x_2, x_3, \ldots, x_n$ with the corresponding weights $w_1, w_2, w_3, \ldots, w_n$ is computed by:

WEIGHTED MEAN

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots + w_n} \qquad (3\text{–}3)$$

This may be shortened to:

$$\bar{x}_w = \frac{\Sigma(wx)}{\Sigma w}$$

Note that the denominator of a weighted mean is always the sum of the weights.

EXAMPLE

The Carter Construction Company pays its hourly employees $16.50, $19.00, or $25.00 per hour. There are 26 hourly employees, 14 of whom are paid at the $16.50 rate, 10 at the $19.00 rate, and 2 at the $25.00 rate. What is the mean hourly rate paid the 26 employees?

SOLUTION

To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate. From formula (3–3), the mean hourly rate is

$$\bar{x}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

The weighted mean hourly wage is rounded to $18.12.

SELF-REVIEW 3–4

Springers sold 95 Antonelli men's suits for the regular price of $400. For the spring sale, the suits were reduced to $200 and 126 were sold. At the final clearance, the price was reduced to $100 and the remaining 79 suits were sold.

(a) What was the weighted mean price of an Antonelli suit?
(b) Springers paid $200 a suit for the 300 suits. Comment on the store's profit per suit if a salesperson receives a $25 commission for each one sold.

EXERCISES

23. In June, an investor purchased 300 shares of Oracle (an information technology company) stock at $20 per share. In August, she purchased an additional 400 shares at $25 per share. In November, she purchased an additional 400 shares, but the stock declined to $23 per share. What is the weighted mean price per share?

24. The Bookstall Inc. is a specialty bookstore concentrating on used books sold via the Internet. Paperbacks are $1.00 each, and hardcover books are $3.50. Of the 50 books sold last Tuesday morning, 40 were paperback and the rest were hardcover. What was the weighted mean price of a book?

25. The Loris Healthcare System employs 200 persons on the nursing staff. Fifty are nurse's aides, 50 are practical nurses, and 100 are registered nurses. Nurse's aides receive $8 an hour, practical nurses $15 an hour, and registered nurses $24 an hour. What is the weighted mean hourly wage?

26. Andrews and Associates specialize in corporate law. They charge $100 an hour for researching a case, $75 an hour for consultations, and $200 an hour for writing a brief. Last week one of the associates spent 10 hours consulting with her client, 10 hours researching the case, and 20 hours writing the brief. What was the weighted mean hourly charge for her legal services?

LO3-3
Compute and interpret the geometric mean.

THE GEOMETRIC MEAN

The geometric mean is useful in finding the average change of percentages, ratios, indexes, or growth rates over time. It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the gross domestic product, which compound or build on each other. The geometric mean of a set of n positive numbers is defined as the nth root of the product of n values. The formula for the geometric mean is written:

GEOMETRIC MEAN $$GM = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} \quad \textbf{(3–4)}$$

The geometric mean will always be less than or equal to (never more than) the arithmetic mean. Also, all the data values must be positive.

As an example of the geometric mean, suppose you receive a 5% increase in salary this year and a 15% increase next year. The average annual percent increase is 9.886%,

not 10.0%. Why is this so? We begin by calculating the geometric mean. Recall, for example, that a 5% increase in salary is 105%. We will write it as 1.05.

$$GM = \sqrt{(1.05)(1.15)} = 1.09886$$

This can be verified by assuming that your monthly earning was $3,000 to start and you received two increases of 5% and 15%.

$$\begin{aligned} \text{Raise 1} &= \$3{,}000(.05) = \$150.00 \\ \text{Raise 2} &= \$3{,}150(.15) = \underline{\ 472.50} \\ \text{Total} & \qquad\qquad\quad \$622.50 \end{aligned}$$

Your total salary increase is $622.50. This is equivalent to:

$$\begin{aligned} \$3{,}000.00(.09886) &= \$296.59 \\ \$3{,}296.58(.09886) &= \underline{\ 325.91} \\ & \qquad\quad \$622.50 \end{aligned}$$

The following example shows the geometric mean of several percentages.

EXAMPLE

The return on investment earned by Atkins Construction Company for four successive years was 30%, 20%, −40%, and 200%. What is the geometric mean rate of return on investment?

SOLUTION

The number 1.3 represents the 30% return on investment, which is the "original" investment of 1.0 plus the "return" of 0.3. The number 0.6 represents the loss of 40%, which is the original investment of 1.0 less the loss of 0.4. This calculation assumes the total return each period is reinvested or becomes the base for the next period. In other words, the base for the second period is 1.3 and the base for the third period is (1.3)(1.2) and so forth.

Then the geometric mean rate of return is 29.4%, found by

$$GM = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} = \sqrt[4]{(1.3)(1.2)(0.6)(3.0)} = \sqrt[4]{2.808} = 1.294$$

The geometric mean is the fourth root of 2.808. So, the average rate of return (compound annual growth rate) is 29.4%.

Notice also that if you compute the arithmetic mean [(30 + 20 − 40 + 200)/4 = 52.5], you would have a much larger number, which would overstate the true rate of return!

A second application of the geometric mean is to find an average percentage change over a period of time. For example, if you earned $45,000 in 2004 and $100,000 in 2016, what is your annual rate of increase over the period? It is 6.88%. The rate of increase is determined from the following formula.

RATE OF INCREASE OVER TIME

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at start of period}}} - 1 \qquad \textbf{(3–5)}$$

In formula 3-5 above, n is the number of periods. An example will show the details of finding the average annual percent increase.

EXAMPLE

During the decade of the 1990s, and into the 2000s, Las Vegas, Nevada, was one of the fastest-growing cities in the United States. The population increased from 258,295 in 1990 to 613,599 in 2014. This is an increase of 355,304 people, or a 137.56% increase over the period. The population has more than doubled. What is the average *annual* percent increase?

SOLUTION

There are 24 years between 1990 and 2014, so $n = 24$. Then the geometric mean formula (3–5) as applied to this problem is:

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at start of period}}} - 1.0 = \sqrt[24]{\frac{613{,}599}{258{,}295}} - 1.0 = 1.0367 - 1.0 = .0367$$

To summarize, the steps to compute the geometric mean are:

1. Divide the value at the end of the period by the value at the beginning of the period.
2. Find the *n*th root of the ratio, where *n* is the number of periods.
3. Subtract one.

The value of .0367 indicates that the average annual growth over the period was 3.67%. To put it another way, the population of Las Vegas increased at a rate of 3.67% per year from 1990 to 2014.

SELF-REVIEW 3–5

1. The percent increase in sales for the last 4 years at Combs Cosmetics were 4.91, 5.75, 8.12, and 21.60.
   (a) Find the geometric mean percent increase.
   (b) Find the arithmetic mean percent increase.
   (c) Is the arithmetic mean equal to or greater than the geometric mean?
2. Production of Cablos trucks increased from 23,000 units in 1996 to 120,520 in 2016. Find the geometric mean annual percent increase.

EXERCISES

27. Compute the geometric mean of the following monthly percent increases: 8, 12, 14, 26, and 5.

28. Compute the geometric mean of the following weekly percent increases: 2, 8, 6, 4, 10, 6, 8, and 4.

29. Listed below is the percent increase in sales for the MG Corporation over the last 5 years. Determine the geometric mean percent increase in sales over the period.

9.4	13.8	11.7	11.9	14.7

30. In 2001, a total of 40,244,000 taxpayers in the United States filed their individual tax returns electronically. By the year 2015, the number increased to 128,653,000. What is the geometric mean annual increase for the period?

31. The Consumer Price Index is reported monthly by the U.S. Bureau of Labor Statistics. It reports the change in prices for a market basket of goods from one period to another. The index for 2000 was 172.2. By 2015, it increased to 236.525. What was the geometric mean annual increase for the period?

32. JetBlue Airways is an American low-cost airline headquartered in New York City. Its main base is John F. Kennedy International Airport. JetBlue's revenue in 2002 was $635.2 million. By 2014, revenue had increased to $5,817.0 million. What was the geometric mean annual increase for the period?

33. In 2000, there were 720,000 cell phone subscribers worldwide. By 2015, the number of cell phone subscribers increased to 752,000,000. What is the geometric mean annual increase for the period?

34. The information below shows the cost for a year of college in public and private colleges in 2002–03 and 2015–16. What is the geometric mean annual increase for the period for the two types of colleges? Compare the rates of increase.

Type of College	2002–03	2015–16
Public	$ 4,960	$23,893
Private	18,056	32,405

LO3-4
Compute and interpret the range, variance, and standard deviation.

WHY STUDY DISPERSION?

A measure of location, such as the mean, median, or mode, only describes the center of the data. It is valuable from that standpoint, but it does not tell us anything about the spread of the data. For example, if your nature guide told you that the river ahead averaged 3 feet in depth, would you want to wade across on foot without additional information? Probably not. You would want to know something about the variation in the depth. Is the maximum depth of the river 3.25 feet and the minimum 2.75 feet? If that is the case, you would probably agree to cross. What if you learned the river depth ranged from 0.50 foot to 5.5 feet? Your decision would probably be not to cross. Before making a decision about crossing the river, you want information on both the typical depth and the dispersion in the depth of the river.

A small value for a measure of dispersion indicates that the data are clustered closely, say, around the arithmetic mean. The mean is therefore considered representative of the data. Conversely, a large measure of dispersion indicates that the mean is not reliable. Refer to Chart 3–5. The 100 employees of Hammond Iron Works Inc., a steel fabricating company, are organized into a histogram based on the number of years of employment with the company. The mean is 4.9 years, but the spread of the data is from 6 months to 16.8 years. The mean of 4.9 years is not very representative of all the employees.

STATISTICS IN ACTION

The U.S. Postal Service has tried to become more "user friendly" in the last several years. A recent survey showed that customers were interested in more *consistency* in the time it takes to make a delivery. Under the old conditions, a local letter might take only one day to deliver, or it might take several. "Just tell me how many days ahead I need to mail the birthday card to Mom so it gets there on her birthday, not early, not late," was a common complaint. The level of consistency is measured by the standard deviation of the delivery times.

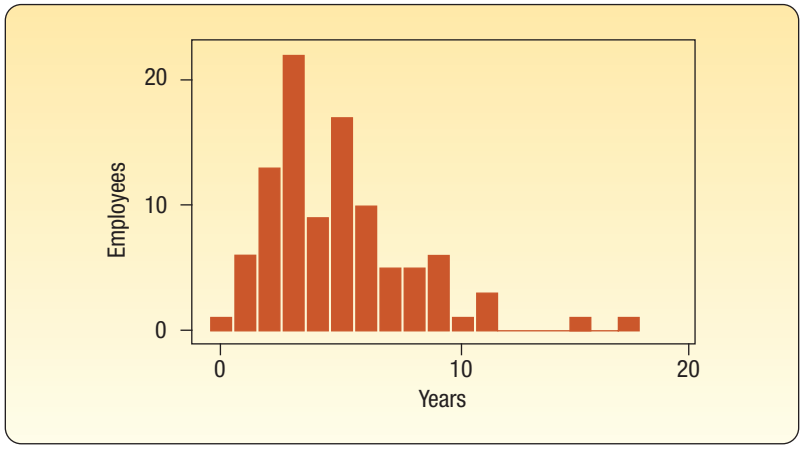


CHART 3–5 Histogram of Years of Employment at Hammond Iron Works Inc.

A second reason for studying the dispersion in a set of data is to compare the spread in two or more distributions. Suppose, for example, that the new Vision Quest LCD computer monitor is assembled in Baton Rouge and also in Tucson. The arithmetic mean hourly output in both the Baton Rouge plant and the Tucson plant is 50. Based on

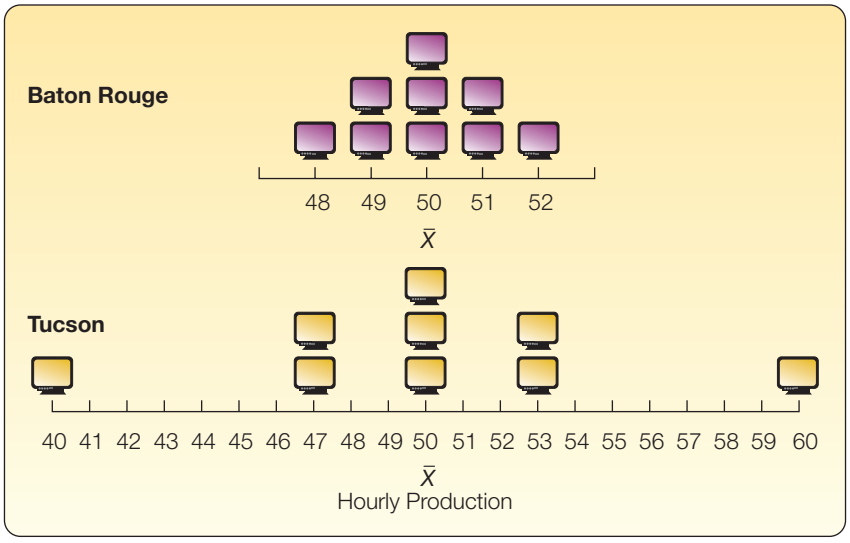


CHART 3–6 Hourly Production of Computer Monitors at the Baton Rouge and Tucson Plants

the two means, you might conclude that the distributions of the hourly outputs are identical. Production records for 9 hours at the two plants, however, reveal that this conclusion is not correct (see Chart 3–6). Baton Rouge production varies from 48 to 52 assemblies per hour. Production at the Tucson plant is more erratic, ranging from 40 to 60 per hour. Therefore, the hourly output for Baton Rouge is clustered near the mean of 50; the hourly output for Tucson is more dispersed.

We will consider several measures of dispersion. The range is based on the maximum and minimum values in the data set; that is, only two values are considered. The variance and the standard deviation use all the values in a data set and are based on deviations from the arithmetic mean.

Range

The simplest measure of dispersion is the **range.** It is the difference between the maximum and minimum values in a data set. In the form of an equation:

RANGE $\text{Range} = \text{Maximum value} - \text{Minimum value}$ **(3–6)**

The range is widely used in production management and control applications because it is very easy to calculate and understand.

EXAMPLE

Refer to Chart 3–6 above. Find the range in the number of computer monitors produced per hour for the Baton Rouge and the Tucson plants. Interpret the two ranges.

SOLUTION

The range of the hourly production of computer monitors at the Baton Rouge plant is 4, found by the difference between the maximum hourly production of 52 and

the minimum of 48. The range in the hourly production for the Tucson plant is 20 computer monitors, found by 60 – 40. We therefore conclude that (1) there is less dispersion in the hourly production in the Baton Rouge plant than in the Tucson plant because the range of 4 computer monitors is less than a range of 20 computer monitors and (2) the production is clustered more closely around the mean of 50 at the Baton Rouge plant than at the Tucson plant (because a range of 4 is less than a range of 20). Thus, the mean production in the Baton Rouge plant (50 computer monitors) is a more representative measure of location than the mean of 50 computer monitors for the Tucson plant.

Variance

A limitation of the range is that it is based on only two values, the maximum and the minimum; it does not take into consideration all of the values. The **variance** does. It measures the mean amount by which the values in a population, or sample, vary from their mean. In terms of a definition:

VARIANCE The arithmetic mean of the squared deviations from the mean.

The following example illustrates how the variance is used to measure dispersion.

EXAMPLE

© Sorbis/Shutterstock.com

The chart below shows the number of cappuccinos sold at the Starbucks in the Orange County airport and the Ontario, California, airport between 4 and 5 p.m. for a sample of 5 days last month.

California Airports	
Orange County	Ontario
20	20
40	45
50	50
60	55
80	80

Determine the mean, median, range, and variance for each location. Comment on the similarities and differences in these measures.

SOLUTION

The mean, median, and range for each of the airport locations are reported as part of an Excel spreadsheet.

	A	B	C
1		California Airports	
2		Orange County	Ontario
3		20	20
4		40	45
5		50	50
6		60	55
7		80	80
8			
9	Mean	50	50
10	Median	50	50
11	Range	60	60

Notice that all three of the measures are exactly the same. Does this indicate that there is no difference in the two sets of data? We get a clearer picture if we calculate the variance. First, for Orange County:

F	G	H
Calculation of Variance for Orange County		
Number Sold	Each Value - Mean	Squared Deviation
20	20 - 50 = -30	900
40	40 - 50 = -10	100
50	50 - 50 = 0	0
60	60 - 50 = 10	100
80	80 - 50 = 30	900
	Total	2000

$$\text{Variance} = \frac{\Sigma(x-\mu)^2}{N} = \frac{(-30^2) + (-10^2) + 0^2 + 10^2 + 30^2}{5} = \frac{2{,}000}{5} = 400$$

The variance is 400. That is, the average squared deviation from the mean is 400.

The following shows the detail of determining the variance for the number of cappuccinos sold at the Ontario Airport.

Calculation of Variance for Ontario		
Number Sold	Each Value - Mean	Squared Deviation
20	20 - 50 = -30	900
45	45 - 50 = -5	25
50	50 - 50 = 0	0
55	55 - 50 = 5	25
80	80 - 50 = 30	900
	Total	1850

$$\text{Variance} = \frac{\Sigma(x-\mu)^2}{N} = \frac{(-30^2) + (-5^2) + 0^2 + 5^2 + 30^2}{5} = \frac{1{,}850}{5} = 370$$

So the mean, median, and range of the cappuccinos sold are the same at the two airports, but the variances are different. The variance at Orange County is 400, but it is 370 at Ontario.

Let's interpret and compare the results of our measures for the two Starbucks airport locations. The mean and median of the two locations are exactly the same, 50 cappuccinos sold. These measures of location suggest the two distributions are the same. The range for both locations is also the same, 60. However, recall that

the range provides limited information about the dispersion because it is based on only two values, the minimum and maximum.

The variances are not the same for the two airports. The variance is based on the differences between each observation and the arithmetic mean. It shows the closeness or clustering of the data relative to the mean or center of the distribution. Compare the variance for Orange County of 400 to the variance for Ontario of 370. Based on the variance, we conclude that the dispersion for the sales distribution of the Ontario Starbucks is more concentrated—that is, nearer the mean of 50—than for the Orange County location.

The variance has an important advantage over the range. It uses all the values in the computation. Recall that the range uses only the highest and the lowest values.

SELF-REVIEW 3–6

The weights of containers being shipped to Ireland are (in thousands of pounds):

95	103	105	110	104	105	112	90

(a) What is the range of the weights?
(b) Compute the arithmetic mean weight.
(c) Compute the variance of the weights.

EXERCISES

For Exercises 35–38, calculate the (a) range, (b) arithmetic mean, (c) variance, and (d) interpret the statistics.

35. FILE During last weekend's sale, there were five customer service representatives on duty at the Electronic Super Store. The numbers of HDTVs these representatives sold were 5, 8, 4, 10, and 3.

36. FILE The Department of Statistics at Western State University offers eight sections of basic statistics. Following are the numbers of students enrolled in these sections: 34, 46, 52, 29, 41, 38, 36, and 28.

37. FILE Dave's Automatic Door installs automatic garage door openers. The following list indicates the number of minutes needed to install 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.

38. FILE All eight companies in the aerospace industry were surveyed as to their return on investment last year. The results are: 10.6%, 12.6%, 14.8%, 18.2%, 12.0%, 14.8%, 12.2%, and 15.6%.

39. FILE Ten young adults living in California rated the taste of a newly developed sushi pizza topped with tuna, rice, and kelp on a scale of 1 to 50, with 1 indicating they did not like the taste and 50 that they did. The ratings were:

34	39	40	46	33	31	34	14	15	45

In a parallel study, 10 young adults in Iowa rated the taste of the same pizza. The ratings were:

28	25	35	16	25	29	24	26	17	20

As a market researcher, compare the potential for sushi pizza in the two markets.

40. FILE The personnel files of all eight employees at the Pawnee location of Acme Carpet Cleaners Inc. revealed that during the last 6-month period they lost the following number of days due to illness:

2	0	6	3	10	4	1	2

All eight employees during the same period at the Chickpee location of Acme Carpets revealed they lost the following number of days due to illness:

2 0 1 0 5 0 1 0

As the director of human resources, compare the two locations. What would you recommend?

Population Variance

In the previous example, we developed the concept of variance as a measure of dispersion. Similar to the mean, we can calculate the variance of a population or the variance of a sample. The formula to compute the population variance is:

POPULATION VARIANCE

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} \qquad (3\text{–}7)$$

where:

- σ^2 is the population variance (σ is the lowercase Greek letter sigma). It is read as "sigma squared."
- x is the value of a particular observation in the population.
- μ is the arithmetic mean of the population.
- N is the number of observations in the population.

The process for computing the variance is implied by the formula.

1. Begin by finding the mean.
2. Find the difference between each observation and the mean, and square that difference.
3. Sum all the squared differences.
4. Divide the sum of the squared differences by the number of items in the population.

So the population variance is the mean of the squared difference between each value and the mean. For populations whose values are near the mean, the variance will be small. For populations whose values are dispersed from the mean, the population variance will be large.

The variance overcomes the weakness of the range by using all the values in the population, whereas the range uses only the maximum and minimum values. We overcome the issue where $\Sigma(x - \mu) = 0$ by squaring the differences. Squaring the differences will always result in nonnegative values. The following is another example that illustrates the calculation and interpretation of the variance.

EXAMPLE

The number of traffic citations issued last year by month in Beaufort County, South Carolina, is reported below.

Citations by Month										
January	February	March	April	May	June	July	August	September	October	November	December
19	17	22	18	28	34	45	39	38	44	34	10

Determine the population variance.

SOLUTION

Because we are studying all the citations for a year, the data comprise a population. To determine the population variance, we use formula (3–7). The table below details the calculations.

Month	Citations (x)	$x - \mu$	$(x - \mu)^2$
January	19	−10	100
February	17	−12	144
March	22	−7	49
April	18	−11	121
May	28	−1	1
June	34	5	25
July	45	16	256
August	39	10	100
September	38	9	81
October	44	15	225
November	34	5	25
December	10	−19	361
Total	348	0	1,488

1. We begin by determining the arithmetic mean of the population. The total number of citations issued for the year is 348, so the mean number issued per month is 29.
$$\mu = \frac{\Sigma x}{N} = \frac{19 + 17 + \cdots + 10}{12} = \frac{348}{12} = 29$$
2. Next we find the difference between each observation and the mean. This is shown in the third column of the table. Recall on page 55 in this chapter, the Verizon example showed that the sum of the differences between each value and the mean is 0. This principle is repeated here. The sum of the differences between the mean and the number of citations each month is 0.
3. The next step is to square the difference for each month. That is shown in the fourth column of the table. All the squared differences will be positive. Note that squaring a negative value, or multiplying a negative value by itself, always results in a positive value.
4. The squared differences are totaled. The total of the fourth column is 1,488. That is the term $\Sigma(x - \mu)^2$.
5. Finally, we divide the squared differences by N, the number of observations in the population.
$$\sigma^2 = \frac{\Sigma(x - \sigma)^2}{N} = \frac{1{,}488}{12} = 124$$
So, the population variance for the number of citations is 124.

Like the range, the variance can be used to compare the dispersion in two or more sets of observations. For example, the variance for the number of citations issued in Beaufort County was just computed to be 124. If the variance in the number of citations issued in Marlboro County, South Carolina, is 342.9, we conclude that (1) there is less dispersion in the distribution of the number of citations issued in Beaufort County than in Marlboro County (because 124 is less than 342.9) and (2) the number of citations in Beaufort County is more closely clustered around the mean of 29 than for the number of citations issued in Marlboro County. Thus the mean number of citations issued in Beaufort County is a more representative measure of location than the mean number of citations in Marlboro County.

Population Standard Deviation

When we compute the variance, it is important to understand the unit of measure and what happens when the differences in the numerator are squared. That is, in the previous example, the number of monthly citations is the variable. When we calculate the variance, the unit of measure for the variance is citations squared. Using "squared citations" as a unit of measure is cumbersome.

There is a way out of this difficulty. By taking the square root of the population variance, we can transform it to the same unit of measurement used for the original data. The square root of 124 citations squared is 11.14 citations. The units are now simply citations. The square root of the population variance is the **population standard deviation.**

POPULATION STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}} \qquad \textbf{(3–8)}$$

SELF-REVIEW 3–7

The Philadelphia office of PricewaterhouseCoopers hired five accounting trainees this year. Their monthly starting salaries were $3,536; $3,173; $3,448; $3,121; and $3,622.

(a) Compute the population mean.
(b) Compute the population variance.
(c) Compute the population standard deviation.
(d) The Pittsburgh office hired six trainees. Their mean monthly salary was $3,550, and the standard deviation was $250. Compare the two groups.

EXERCISES

41. Consider these five values a population: 8, 3, 7, 3, and 4.
- **a.** Determine the mean of the population.
- **b.** Determine the variance.

42. Consider these six values a population: 13, 3, 8, 10, 8, and 6.
- **a.** Determine the mean of the population.
- **b.** Determine the variance.

43. The annual report of Dennis Industries cited these primary earnings per common share for the past 5 years: $2.68, $1.03, $2.26, $4.30, and $3.58. If we assume these are population values, what is:
- **a.** The arithmetic mean primary earnings per share of common stock?
- **b.** The variance?

44. Referring to Exercise 43, the annual report of Dennis Industries also gave these returns on stockholder equity for the same 5-year period (in percent): 13.2, 5.0, 10.2, 17.5, and 12.9.
- **a.** What is the arithmetic mean return?
- **b.** What is the variance?

45. Plywood Inc. reported these returns on stockholder equity for the past 5 years: 4.3, 4.9, 7.2, 6.7, and 11.6. Consider these as population values.
- **a.** Compute the range, the arithmetic mean, the variance, and the standard deviation.
- **b.** Compare the return on stockholder equity for Plywood Inc. with that for Dennis Industries cited in Exercise 44.

46. The annual incomes of the five vice presidents of TMV Industries are $125,000; $128,000; $122,000; $133,000; and $140,000. Consider this a population.
- **a.** What is the range?
- **b.** What is the arithmetic mean income?
- **c.** What is the population variance? The standard deviation?
- **d.** The annual incomes of officers of another firm similar to TMV Industries were also studied. The mean was $129,000 and the standard deviation $8,612. Compare the means and dispersions in the two firms.

Sample Variance and Standard Deviation

The formula for the population mean is $\mu = \Sigma x/N$. We just changed the symbols for the sample mean; that is, $\bar{x} = \Sigma x/n$. Unfortunately, the conversion from the population variance to the sample variance is not as direct. It requires a change in the denominator. Instead of substituting n (number in the sample) for N (number in the population), the denominator is $n - 1$. Thus the formula for the **sample variance** is:

SAMPLE VARIANCE

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \qquad (3\text{–}9)$$

where:

- s^2 is the sample variance.
- x is the value of each observation in the sample.
- $\bar{x}$ is the mean of the sample.
- n is the number of observations in the sample.

Why is this change made in the denominator? Although the use of n is logical since $\bar{x}$ is used to estimate μ, it tends to underestimate the population variance, σ^2. The use of $(n - 1)$ in the denominator provides the appropriate correction for this tendency. Because the primary use of sample statistics like s^2 is to estimate population parameters like σ^2, $(n - 1)$ is preferred to n in defining the sample variance. We will also use this convention when computing the sample standard deviation.

EXAMPLE

The hourly wages for a sample of part-time employees at Home Depot are $12, $20, $16, $18, and $19. What is the sample variance?

SOLUTION

The sample variance is computed by using formula (3–9).

$$\bar{x} = \frac{\Sigma x}{n} = \frac{\$85}{5} = \$17$$

Hourly Wage (x)	$x - \bar{x}$	$(x - \bar{x})^2$
$12	−$5	25
20	3	9
16	−1	1
18	1	1
19	2	4
$85	0	40

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{40}{5 - 1}$$

$= 10$ in dollars squared

The sample standard deviation is used as an estimator of the population standard deviation. As noted previously, the population standard deviation is the square root of the population variance. Likewise, the *sample standard deviation is the square root of the sample variance*. The sample standard deviation is determined by:

SAMPLE STANDARD DEVIATION

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \qquad (3\text{–}10)$$

EXAMPLE

The sample variance in the previous example involving hourly wages was computed to be 10. What is the sample standard deviation?

SOLUTION

The sample standard deviation is \$3.16, found by $\sqrt{10}$. Note again that the sample variance is in terms of dollars squared, but taking the square root of 10 gives us \$3.16, which is in the same units (dollars) as the original data.

Software Solution

On page 64, we used Excel to determine the mean, median, and mode of profit for the Applewood Auto Group data. You also will note that it lists the sample variance and sample standard deviation. Excel, like most other statistical software, assumes the data are from a sample.

APPLEWOOD AUTO GROUP

	A	B	C	D	E	F	G	H
1	**Age**	**Profit**	**Location**	**Vehicle-Type**	**Previous**		*Profit*	
2	21	\$1,387	Tionesta	Sedan	0			
3	23	\$1,754	Sheffield	SUV	1		Mean	1843.17
4	24	\$1,817	Sheffield	Hybrid	1		Standard Error	47.97
5	25	\$1,040	Sheffield	Compact	0		Median	1882.50
6	26	\$1,273	Kane	Sedan	1		Mode	1915.00
7	27	\$1,529	Sheffield	Sedan	1		Standard Deviation	643.63
8	27	\$3,082	Kane	Truck	0		Sample Variance	414256.61
9	28	\$1,951	Kane	SUV	1		Kurtosis	-0.22
10	28	\$2,692	Tionesta	Compact	0		Skewness	-0.24
11	29	\$1,342	Kane	Sedan	2		Range	2998
12	29	\$1,206	Sheffield	Sedan	0		Minimum	294
13	30	\$443	Kane	Sedan	3		Maximum	3292
14	30	\$1,621	Sheffield	Truck	1		Sum	331770
15	30	\$754	Olean	Sedan	2		Count	180

SELF-REVIEW 3–8

The years of service for a sample of seven employees at a State Farm Insurance claims office in Cleveland, Ohio, are 4, 2, 5, 4, 5, 2, and 6. What is the sample variance? Compute the sample standard deviation.

EXERCISES

For Exercises 47–52, do the following:

a. Compute the sample variance.
b. Determine the sample standard deviation.

47. Consider these values a sample: 7, 2, 6, 2, and 3.

48. The following five values are a sample: 11, 6, 10, 6, and 7.

49. FILE Dave's Automatic Door, referred to in Exercise 37, installs automatic garage door openers. Based on a sample, following are the times, in minutes, required to install 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.

50. FILE The sample of eight companies in the aerospace industry, referred to in Exercise 38, was surveyed as to their return on investment last year. The results are 10.6, 12.6, 14.8, 18.2, 12.0, 14.8, 12.2, and 15.6.

51. FILE The Houston, Texas, Motel Owner Association conducted a survey regarding weekday motel rates in the area. Listed below is the room rate for business-class guests for a sample of 10 motels.

$101	$97	$103	$110	$78	$87	$101	$80	$106	$88

52. FILE A consumer watchdog organization is concerned about credit card debt. A survey of 10 young adults with credit card debt of more than $2,000 showed they paid an average of just over $100 per month against their balances. Listed below are the amounts each young adult paid last month.

$110	$126	$103	$93	$99	$113	$87	$101	$109	$100

LO3-5
Explain and apply Chebyshev's theorem and the Empirical Rule.

INTERPRETATION AND USES OF THE STANDARD DEVIATION

The standard deviation is commonly used as a measure to compare the spread in two or more sets of observations. For example, the standard deviation of the biweekly amounts invested in the Dupree Paint Company profit-sharing plan is computed to be $7.51. Suppose these employees are located in Georgia. If the standard deviation for a group of employees in Texas is $10.47, and the means are about the same, it indicates that the amounts invested by the Georgia employees are not dispersed as much as those in Texas (because $7.51 < $10.47). Since the amounts invested by the Georgia employees are clustered more closely about the mean, the mean for the Georgia employees is a more reliable measure than the mean for the Texas group.

STATISTICS IN ACTION

Most colleges report the "average class size." This information can be misleading because average class size can be found in several ways. If we find the number of students *in each class* at a particular university, the result is the mean number of students per class. If we compile a list of the class sizes for each student and find the mean class size, we might find the mean to be quite different. One school found the mean number of students in each of its 747 classes to be 40. But when

(*continued*)

Chebyshev's Theorem

We have stressed that a small standard deviation for a set of values indicates that these values are located close to the mean. Conversely, a large standard deviation reveals that the observations are widely scattered about the mean. The Russian mathematician P. L. Chebyshev (1821–1894) developed a theorem that allows us to determine the minimum proportion of the values that lie within a specified number of standard deviations of the mean. For example, according to **Chebyshev's theorem,** at least three out of every four, or 75%, of the values must lie between the mean plus two standard deviations and the mean minus two standard deviations. This relationship applies regardless of the shape of the distribution. Further, at least eight of nine values, or 88.9%, will lie between plus three standard deviations and minus three standard deviations of the mean. At least 24 of 25 values, or 96%, will lie between plus and minus five standard deviations of the mean.

Chebyshev's theorem states:

CHEBYSHEV'S THEOREM For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least $1 - 1/k^2$, where k is any value greater than 1.

EXAMPLE

Dupree Paint Company employees contribute a mean of $51.54 to the company's profit-sharing plan every two weeks. The standard deviation of biweekly contributions is $7.51. At least what percent of the contributions lie within plus 3.5 standard deviations and minus 3.5 standard deviations of the mean, that is between $25.26 and $77.83?

SOLUTION

About 92%, found by

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

(*continued from p. 79*)
it found the mean from a list of the class sizes of each student, it was 147. Why the disparity? Because there are few students in the small classes and a larger number of students in the larger classes, which has the effect of increasing the mean class size when it is calculated this way. A school could reduce this mean class size for each student by reducing the number of students in each class. That is, cut out the large freshman lecture classes.

The Empirical Rule

Chebyshev's theorem applies to any set of values; that is, the distribution of values can have any shape. However, for a symmetrical, bell-shaped distribution such as the one in Chart 3–7, we can be more precise in explaining the dispersion about the mean. These relationships involving the standard deviation and the mean are described by the **Empirical Rule,** sometimes called the **Normal Rule.**

EMPIRICAL RULE For a symmetrical, bell-shaped frequency distribution, approximately 68% of the observations will lie within plus and minus one standard deviation of the mean; about 95% of the observations will lie within plus and minus two standard deviations of the mean; and practically all (99.7%) will lie within plus and minus three standard deviations of the mean.

These relationships are portrayed graphically in Chart 3–7 for a bell-shaped distribution with a mean of 100 and a standard deviation of 10.

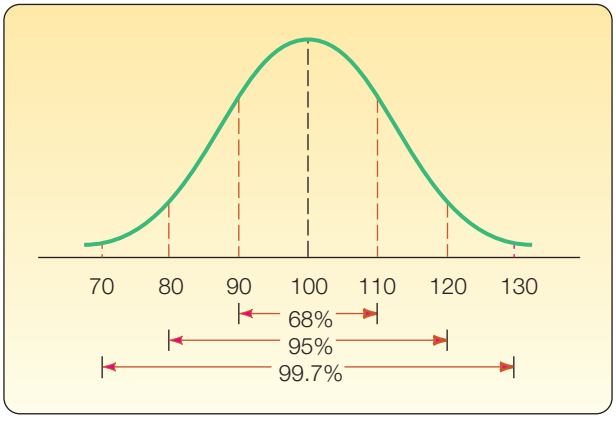


CHART 3–7 A Symmetrical, Bell-Shaped Curve Showing the Relationships between the Standard Deviation and the Percentage of Observations

Applying the Empirical Rule, if a distribution is symmetrical and bell-shaped, practically all of the observations lie between the mean plus and minus three standard deviations. Thus, if $\bar{x} = 100$ and $s = 10$, practically all the observations lie between 100 + 3(10) and 100 – 3(10), or 70 and 130. The estimated range is therefore 60, found by 130 – 70.

Conversely, if we know that the range is 60 and the distribution is bell-shaped, we can approximate the standard deviation by dividing the range by 6. For this illustration: range ÷ 6 = 60 ÷ 6 = 10, the standard deviation.

EXAMPLE

A sample of the rental rates at University Park Apartments approximates a symmetrical, bell-shaped distribution. The sample mean is $500; the standard deviation is $20. Using the Empirical Rule, answer these questions:

1. About 68% of the monthly rentals are between what two amounts?
2. About 95% of the monthly rentals are between what two amounts?
3. Almost all of the monthly rentals are between what two amounts?

SOLUTION

1. About 68% are between $480 and $520, found by $\bar{x} \pm 1s = \$500 \pm 1(\$20)$.
2. About 95% are between $460 and $540, found by $\bar{x} \pm 2s = \$500 \pm 2(\$20)$.
3. Almost all (99.7%) are between $440 and $560, found by $\bar{x} \pm 3s = \$500 \pm 3(\$20)$.

SELF-REVIEW 3–9

The Pitney Pipe Company is one of several domestic manufacturers of PVC pipe. The quality control department sampled 600 10-foot lengths. At a point 1 foot from the end of the pipe, they measured the outside diameter. The mean was 14.0 inches and the standard deviation 0.1 inch.

(a) If we do not know the shape of the distribution of outside pipe diameters, at least what percent of the observations will be between 13.85 inches and 14.15 inches?

(b) If we assume that the distribution of diameters is symmetrical and bell-shaped, about 95% of the observations will be between what two values?

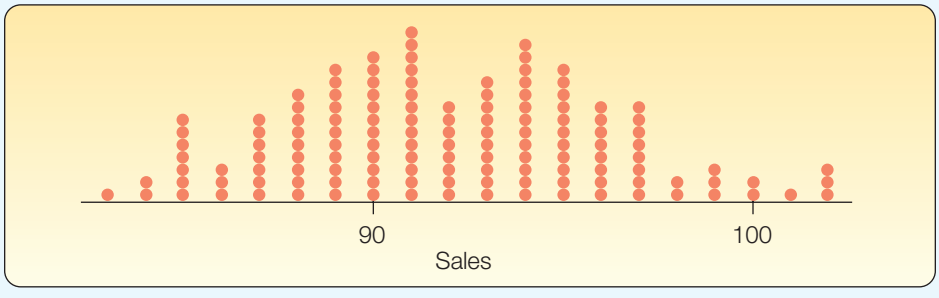
EXERCISES

53. According to Chebyshev's theorem, at least what percent of any set of observations will be within 1.8 standard deviations of the mean?

54. The mean income of a group of sample observations is $500; the standard deviation is $40. According to Chebyshev's theorem, at least what percent of the incomes will lie between $400 and $600?

55. The distribution of the weights of a sample of 1,400 cargo containers is symmetric and bell-shaped. According to the Empirical Rule, what percent of the weights will lie:
   a. Between $\bar{x} - 2s$ and $\bar{x} + 2s$?
   b. Between $\bar{x}$ and $\bar{x} + 2s$? Above $\bar{x} + 2s$?

56. The following graph portrays the distribution of the number of spicy chicken sandwiches sold at a nearby Wendy's for the last 141 days. The mean number of sandwiches sold per day is 91.9 and the standard deviation is 4.67.



If we use the Empirical Rule, sales will be between what two values on 68% of the days? Sales will be between what two values on 95% of the days?

LO3-6
Compute the mean and standard deviation of grouped data.

STATISTICS IN ACTION

During the 2016 Major League Baseball season, DJ LeMahieu of the Colorado Rockies had the highest batting average at .348. Tony Gwynn hit .394 in the strike-shortened season of 1994, and Ted Williams hit .406 in 1941. No one has hit over .400 since 1941. The mean batting average has remained constant at about .260 for more than 100 years, but the standard deviation declined from .049 to .031. This indicates less dispersion in the batting averages today and helps explain the lack of any .400 hitters in recent times.

THE MEAN AND STANDARD DEVIATION OF GROUPED DATA

In most instances, measures of location, such as the mean, and measures of dispersion, such as the standard deviation, are determined by using the individual values. Statistical software packages make it easy to calculate these values, even for large data sets. However, sometimes we are given only the frequency distribution and wish to estimate the mean or standard deviation. In the following discussion, we show how we can estimate the mean and standard deviation from data organized into a frequency distribution. We should stress that a mean or a standard deviation from grouped data is an *estimate* of the corresponding actual values.

Arithmetic Mean of Grouped Data

To approximate the arithmetic mean of data organized into a frequency distribution, we begin by assuming the observations in each class are represented by the *midpoint* of the class. The mean of a sample of data organized in a frequency distribution is computed by:

ARITHMETIC MEAN OF GROUPED DATA

$$\bar{x} = \frac{\Sigma fM}{n} \qquad \textbf{(3–11)}$$

where:

- $\bar{x}$ is the sample mean.
- M is the midpoint of each class.
- f is the frequency in each class.
- fM is the frequency in each class times the midpoint of the class.
- Σfm is the sum of these products.
- n is the total number of frequencies.

EXAMPLE

The computations for the arithmetic mean of data grouped into a frequency distribution will be shown based on the Applewood Auto Group profit data. Recall in Chapter 2, in Table 2–7 on page 30, we constructed a frequency distribution for the vehicle profit. The information is repeated below. Determine the arithmetic mean profit per vehicle.

Profit	Frequency
$ 200 up to $ 600	8
600 up to 1,000	11
1,000 up to 1,400	23
1,400 up to 1,800	38
1,800 up to 2,200	45
2,200 up to 2,600	32
2,600 up to 3,000	19
3,000 up to 3,400	4
Total	180

SOLUTION

The mean vehicle selling price can be estimated from data grouped into a frequency distribution. To find the estimated mean, assume the midpoint of each class is representative of the data values in that class. Recall that the midpoint of a class

is halfway between the lower class limits of two consecutive classes. To find the midpoint of a particular class, we add the lower limits of two consecutive classes and divide by 2. Hence, the midpoint of the first class is $400, found by ($200 + $600)/2. We assume the value of $400 is representative of the eight values in that class. To put it another way, we assume the sum of the eight values in this class is $3,200, found by 8($400). We continue the process of multiplying the class midpoint by the class frequency for each class and then sum these products. The results are summarized in Table 3–1.

TABLE 3–1 Profit on 180 Vehicles Sold Last Month at Applewood Auto Group

Profit	Frequency (*f*)	Midpoint (*M*)	*fM*
$ 200 up to $ 600	8	$ 400	$ 3,200
600 up to 1,000	11	800	8,800
1,000 up to 1,400	23	1,200	27,600
1,400 up to 1,800	38	1,600	60,800
1,800 up to 2,200	45	2,000	90,000
2,200 up to 2,600	32	2,400	76,800
2,600 up to 3,000	19	2,800	53,200
3,000 up to 3,400	4	3,200	12,800
Total	180		$333,200

Solving for the arithmetic mean using formula (3–11), we get:

$$\bar{x} = \frac{\Sigma fM}{n} = \frac{\$333,200}{180} = \$1,851.11$$

We conclude that the mean profit per vehicle is about $1,851.

Standard Deviation of Grouped Data

To calculate the standard deviation of data grouped into a frequency distribution, we need to adjust formula (3–10) slightly. We weight each of the squared differences by the number of frequencies in each class. The formula is:

STANDARD DEVIATION, GROUPED DATA

$$s = \sqrt{\frac{\Sigma f(M - \bar{x})^2}{n - 1}} \qquad \textbf{(3–12)}$$

where:

- s is the sample standard deviation.
- M is the midpoint of the class.
- f is the class frequency.
- n is the number of observations in the sample.
- $\bar{x}$ is the sample mean.

EXAMPLE

Refer to the frequency distribution for the Applewood Auto Group profit data reported in Table 3–1. Compute the standard deviation of the vehicle selling prices.

SOLUTION

Following the same practice used earlier for computing the mean of data grouped into a frequency distribution, f is the class frequency, M the class midpoint, and n the number of observations.

Profit	Frequency (f)	Midpoint (M)	fM	$(M - \bar{x})$	$(M - \bar{x})^2$	$f(M - \bar{x})^2$
$ 200 up to $ 600	8	400	3,200	−1,451	2,105,401	16,843,208
600 up to 1,000	11	800	8,800	−1,051	1,104,601	12,150,611
1,000 up to 1,400	23	1,200	27,600	−651	423,801	9,747,423
1,400 up to 1,800	38	1,600	60,800	−251	63,001	2,394,038
1,800 up to 2,200	45	2,000	90,000	149	22,201	999,045
2,200 up to 2,600	32	2,400	76,800	549	301,401	9,644,832
2,600 up to 3,000	19	2,800	53,200	949	900,601	17,111,419
3,000 up to 3,400	4	3,200	12,800	1,349	1,819,801	7,279,204
Total	180		333,200			76,169,780

To find the standard deviation:

Step 1: Subtract the mean from the class midpoint. That is, find $(M - \bar{x})$ = ($400 − $1,851 = −$1,451) for the first class, for the second class ($800 − $1,851 = −$1,051), and so on.

Step 2: Square the difference between the class midpoint and the mean. For the first class, it would be $(\$400 - \$1,851)^2 = 2,105,401$, for the second class $(\$800 - \$1,851)^2 = 1,104,601$, and so on.

Step 3: Multiply the squared difference between the class midpoint and the mean by the class frequency. For the first class, the value is $8(\$400 - \$1,851)^2 = 16,843,208$; for the second, $11(\$800 - \$1,851)^2 = 12,150,611$, and so on.

Step 4: Sum the $f(M - \bar{x})^2$. The total is 76,169,920. To find the standard deviation, we insert these values in formula (3–12).

$$s = \sqrt{\frac{\Sigma f(M - \bar{x})^2}{n - 1}} = \sqrt{\frac{76,169,780}{180 - 1}} = 652.33$$

The mean and the standard deviation calculated from the data grouped into a frequency distribution are usually close to the values calculated from raw data. The grouped data result in some loss of information. For the vehicle profit example, the mean profit reported in the Excel output on page 64 is $1,843.17 and the standard deviation is $643.63. The respective values estimated from data grouped into a frequency distribution are $1,851.11 and $652.33. The difference in the means is $7.94, or about 0.4%. The standard deviations differ by $8.70, or 1.4%. Based on the percentage difference, the estimates are very close to the actual values.

SELF-REVIEW 3–10

The net incomes of a sample of twenty container shipping companies were organized into the following table:

Net Income ($ millions)	Number of Companies
2 up to 6	1
6 up to 10	4
10 up to 14	10
14 up to 18	3
18 up to 22	2

(a) What is the table called?
(b) Based on the distribution, what is the estimate of the arithmetic mean net income?
(c) Based on the distribution, what is the estimate of the standard deviation?

EXERCISES

57. When we compute the mean of a frequency distribution, why do we refer to this as an *estimated* mean?

58. Estimate the mean and the standard deviation of the following frequency distribution showing the number of times students eat at campus dining places in a month.

Class	Frequency
0 up to 5	2
5 up to 10	7
10 up to 15	12
15 up to 20	6
20 up to 25	3

59. Estimate the mean and the standard deviation of the following frequency distribution showing the ages of the first 60 people in line on Black Friday at a retail store.

Class	Frequency
20 up to 30	7
30 up to 40	12
40 up to 50	21
50 up to 60	18
60 up to 70	12

60. SCCoast, an Internet provider in the Southeast, developed the following frequency distribution on the age of Internet users. Estimate the mean and the standard deviation.

Age (years)	Frequency
10 up to 20	3
20 up to 30	7
30 up to 40	18
40 up to 50	20
50 up to 60	12

61. The IRS was interested in the number of individual tax forms prepared by small accounting firms. The IRS randomly sampled 50 public accounting firms with 10 or fewer employees in the Dallas–Fort Worth area. The following frequency table reports the results of the study. Estimate the mean and the standard deviation.

Number of Clients	Frequency
20 up to 30	1
30 up to 40	15
40 up to 50	22
50 up to 60	8
60 up to 70	4

62. Advertising expenses are a significant component of the cost of goods sold. Listed below is a frequency distribution showing the advertising expenditures for 60 manufacturing companies located in the Southwest. Estimate the mean and the standard deviation of advertising expenses.

Advertising Expenditure ($ millions)	Number of Companies
25 up to 35	5
35 up to 45	10
45 up to 55	21
55 up to 65	16
65 up to 75	8
Total	60

ETHICS AND REPORTING RESULTS

In Chapter 1, we discussed the ethical and unbiased reporting of statistical results. While you are learning about how to organize, summarize, and interpret data using statistics, it is also important to understand statistics so that you can be an intelligent consumer of information.

In this chapter, we learned how to compute numerical descriptive statistics. Specifically, we showed how to compute and interpret measures of location for a data set: the mean, median, and mode. We also discussed the advantages and disadvantages for each statistic. For example, if a real estate developer tells a client that the average home in a particular subdivision sold for $150,000, we assume that $150,000 is a representative selling price for all the homes. But suppose that the client also asks what the median sales price is, and the median is $60,000. Why was the developer only reporting the mean price? This information is extremely important to a person's decision making when buying a home. Knowing the advantages and disadvantages of the mean, median, and mode is important as we report statistics and as we use statistical information to make decisions.

We also learned how to compute measures of dispersion: range, variance, and standard deviation. Each of these statistics also has advantages and disadvantages. Remember that the range provides information about the overall spread of a distribution. However, it does not provide any information about how the data are clustered or concentrated around the center of the distribution. As we learn more about statistics, we need to remember that when we use statistics we must maintain an independent and principled point of view. Any statistical report requires objective and honest communication of the results.

CHAPTER SUMMARY

I. A measure of location is a value used to describe the central tendency of a set of data.

A. The arithmetic mean is the most widely reported measure of location.

1. It is calculated by adding the values of the observations and dividing by the total number of observations.

a. The formula for the population mean of ungrouped or raw data is

$$\mu = \frac{\Sigma x}{N} \qquad \textbf{(3–1)}$$

b. The formula for the sample mean is

$$\bar{x} = \frac{\Sigma x}{n} \qquad \textbf{(3–2)}$$

c. The formula for the sample mean of data in a frequency distribution is

$$\bar{x} = \frac{\Sigma fM}{n} \quad \textbf{(3–11)}$$

2. The major characteristics of the arithmetic mean are:
- **a.** At least the interval scale of measurement is required.
- **b.** All the data values are used in the calculation.
- **c.** A set of data has only one mean. That is, it is unique.
- **d.** The sum of the deviations from the mean equals 0.

B. The median is the value in the middle of a set of ordered data.
1. To find the median, sort the observations from minimum to maximum and identify the middle value.
2. The major characteristics of the median are:
- **a.** At least the ordinal scale of measurement is required.
- **b.** It is not influenced by extreme values.
- **c.** Fifty percent of the observations are larger than the median.
- **d.** It is unique to a set of data.

C. The mode is the value that occurs most often in a set of data.
1. The mode can be found for nominal-level data.
2. A set of data can have more than one mode.

D. The weighted mean is found by multiplying each observation by its corresponding weight.
1. The formula for determining the weighted mean is

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots + w_n} \quad \textbf{(3–3)}$$

E. The geometric mean is the *n*th root of the product of *n* positive values.
1. The formula for the geometric mean is

$$GM = \sqrt[n]{(x_1)(x_2)(x_3)\cdots(x_n)} \quad \textbf{(3–4)}$$

2. The geometric mean is also used to find the rate of change from one period to another.

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at beginning of period}}} - 1 \quad \textbf{(3–5)}$$

3. The geometric mean is always equal to or less than the arithmetic mean.

II. The dispersion is the variation or spread in a set of data.

A. The range is the difference between the maximum and minimum values in a set of data.
1. The formula for the range is

$$\text{Range} = \text{Maximum value} - \text{Minimum value} \quad \textbf{(3–6)}$$

2. The major characteristics of the range are:
- **a.** Only two values are used in its calculation.
- **b.** It is influenced by extreme values.
- **c.** It is easy to compute and to understand.

B. The variance is the mean of the squared deviations from the arithmetic mean.
1. The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} \quad \textbf{(3–7)}$$

2. The formula for the sample variance is

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \quad \textbf{(3–9)}$$

3. The major characteristics of the variance are:
- **a.** All observations are used in the calculation.
- **b.** The units are somewhat difficult to work with; they are the original units squared.

C. The standard deviation is the square root of the variance.
1. The major characteristics of the standard deviation are:
- **a.** It is in the same units as the original data.
- **b.** It is the square root of the average squared distance from the mean.
- **c.** It cannot be negative.
- **d.** It is the most widely reported measure of dispersion.

2. The formula for the sample standard deviation is

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} \qquad \textbf{(3–10)}$$

3. The formula for the standard deviation of grouped data is

$$s = \sqrt{\frac{\Sigma f(M - \bar{x})^2}{n - 1}} \qquad \textbf{(3–12)}$$

III. We use the standard deviation to describe a frequency distribution by applying Chebyshev's theorem or the Empirical Rule.

A. Chebyshev's theorem states that regardless of the shape of the distribution, at least $1 - 1/k^2$ of the observations will be within k standard deviations of the mean, where k is greater than 1.

B. The Empirical Rule states that for a bell-shaped distribution about 68% of the values will be within one standard deviation of the mean, 95% within two, and virtually all within three.

PRONUNCIATION KEY

SYMBOL	MEANING	PRONUNCIATION
μ	Population mean	*mu*
Σ	Operation of adding	*sigma*
Σx	Adding a group of values	*sigma x*
$\bar{x}$	Sample mean	*x bar*
$\bar{x}_w$	Weighted mean	*x bar sub w*
GM	Geometric mean	*G M*
ΣfM	Adding the product of the frequencies and the class midpoints	*sigma f M*
σ^2	Population variance	*sigma squared*
σ	Population standard deviation	*sigma*

CHAPTER EXERCISES

63. The accounting firm of Crawford and Associates has five senior partners. Yesterday the senior partners saw six, four, three, seven, and five clients, respectively.

a. Compute the mean and median number of clients seen by the partners.

b. Is the mean a sample mean or a population mean?

c. Verify that $\Sigma(x - \mu) = 0$.

64. Owens Orchards sells apples in a large bag by weight. A sample of seven bags contained the following numbers of apples: 23, 19, 26, 17, 21, 24, 22.

a. Compute the mean and median number of apples in a bag.

b. Verify that $\Sigma(x - \bar{x}) = 0$.

65. FILE A sample of households that subscribe to United Bell Phone Company for landline phone service revealed the following number of calls received per household last week. Determine the mean and the median number of calls received.

52	43	30	38	30	42	12	46	39	37
34	46	32	18	41	5				

66. FILE The Citizens Banking Company is studying the number of times the ATM located in a Loblaws Supermarket at the foot of Market Street is used per day. Following are the number of times the machine was used daily over each of the last 30 days. Determine the mean number of times the machine was used per day.

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

67. FILE A recent study of the laundry habits of Americans included the time in minutes of the wash cycle. A sample of 40 observations follows. Determine the mean and the median of a typical wash cycle.

35	37	28	37	33	38	37	32	28	29
39	33	32	37	33	35	36	44	36	34
40	38	46	39	37	39	34	39	31	33
37	35	39	38	37	32	43	31	31	35

68. FILE Trudy Green works for the True-Green Lawn Company. Her job is to solicit lawn-care business via the telephone. Listed below is the number of appointments she made in each of the last 25 hours of calling. What is the arithmetic mean number of appointments she made per hour? What is the median number of appointments per hour? Write a brief report summarizing the findings.

9	5	2	6	5	6	4	4	7	2	3	6	3
4	4	7	8	4	4	5	5	4	8	3	3	

69. The Split-A-Rail Fence Company sells three types of fence to homeowners in suburban Seattle, Washington. Grade A costs $5.00 per running foot to install, Grade B costs $6.50 per running foot, and Grade C, the premium quality, costs $8.00 per running foot. Yesterday, Split-A-Rail installed 270 feet of Grade A, 300 feet of Grade B, and 100 feet of Grade C. What was the mean cost per foot of fence installed?

70. Rolland Poust is a sophomore in the College of Business at Scandia Tech. Last semester he took courses in statistics and accounting, 3 hours each, and earned an A in both. He earned a B in a 5-hour history course and a B in a 2-hour history of jazz course. In addition, he took a 1-hour course dealing with the rules of basketball so he could get his license to officiate high school basketball games. He got an A in this course. What was his GPA for the semester? Assume that he receives 4 points for an A, 3 for a B, and so on. What measure of central tendency did you calculate? What method did you use?

71. The table below shows the percent of the labor force that is unemployed and the size of the labor force for three counties in northwest Ohio. Jon Elsas is the Regional Director of Economic Development. He must present a report to several companies that are considering locating in northwest Ohio. What would be an appropriate unemployment rate to show for the entire region?

County	Percent Unemployed	Size of Workforce
Wood	4.5	15,300
Ottawa	3.0	10,400
Lucas	10.2	150,600

72. FILE The American Diabetes Association recommends a blood glucose reading of less than 130 for those with Type 2 diabetes. Blood glucose measures the amount of sugar in the blood. Below are the readings for February for a person recently diagnosed with Type 2 diabetes.

112	122	116	103	112	96	115	98	106	111
106	124	116	127	116	108	112	112	121	115
124	116	107	118	123	109	109	106		

a. What is the arithmetic mean glucose reading?
b. What is the median glucose reading?
c. What is the modal glucose reading?

73. The first Super Bowl was played in 1967. The cost for a 30-second commercial was $42,000. The cost of a 30-second commercial for Super Bowl 50 was $4.6 million. What was the geometric mean rate of increase for the 50 year period?

74. A recent article suggested that, if you earn $25,000 a year today and the inflation rate continues at 3% per year, you'll need to make $33,598 in 10 years to have the same buying power. You would need to make $44,771 if the inflation rate jumped to 6%. Confirm that these statements are accurate by finding the geometric mean rate of increase.

75. The ages of a sample of Canadian tourists flying from Toronto to Hong Kong were 32, 21, 60, 47, 54, 17, 72, 55, 33, and 41.
- **a.** Compute the range.
- **b.** Compute the standard deviation.

76. The weights (in pounds) of a sample of five boxes being sent by UPS are 12, 6, 7, 3, and 10.
- **a.** Compute the range.
- **b.** Compute the standard deviation.

77. FILE The enrollments of the 13 public universities in the state of Ohio are listed below.

College	Enrollment
University of Akron	26,106
Bowling Green State University	18,864
Central State University	1,718
University of Cincinnati	44,354
Cleveland State University	17,194
Kent State University	41,444
Miami University	23,902
Ohio State University	62,278
Ohio University	36,493
Shawnee State University	4,230
University of Toledo	20,595
Wright State University	17,460
Youngstown State University	12,512

- **a.** Is this a sample or a population?
- **b.** What is the mean enrollment?
- **c.** What is the median enrollment?
- **d.** What is the range of the enrollments?
- **e.** Compute the standard deviation.

78. Health issues are a concern of managers, especially as they evaluate the cost of medical insurance. A recent survey of 150 executives at Elvers Industries, a large insurance and financial firm located in the Southwest, reported the number of pounds by which the executives were overweight. Compute the mean and the standard deviation.

Pounds Overweight	Frequency
0 up to 6	14
6 up to 12	42
12 up to 18	58
18 up to 24	28
24 up to 30	8

79. FILE The Apollo space program lasted from 1967 until 1972 and included 13 missions. The missions lasted from as little as 7 hours to as long as 301 hours. The duration of each flight is listed below.

9	195	241	301	216	260	7	244	192	147
10	295	142							

- **a.** Explain why the flight times are a population.
- **b.** Find the mean and median of the flight times.
- **c.** Find the range and the standard deviation of the flight times.

80. FILE Creek Ratz is a very popular restaurant located along the coast of northern Florida. They serve a variety of steak and seafood dinners. During the summer beach season, they do not take reservations or accept "call ahead" seating. Management of the restaurant is concerned with the time a patron must wait before being seated for dinner. Listed below is the wait time, in minutes, for the 25 tables seated last Saturday night.

28	39	23	67	37	28	56	40	28	50
51	45	44	65	61	27	24	61	34	44
64	25	24	27	29					

a. Explain why the times are a population.
b. Find the mean and median of the times.
c. Find the range and the standard deviation of the times.

81. FILE A sample of 25 undergraduates reported the following dollar amounts of entertainment expenses last year:

684	710	688	711	722	698	723	743	738	722	696	721	685
763	681	731	736	771	693	701	737	717	752	710	697	

a. Find the mean, median, and mode of this information.
b. What are the range and standard deviation?
c. Use the Empirical Rule to establish an interval that includes about 95% of the observations.

82. FILE The Kentucky Derby is held the first Saturday in May at Churchill Downs in Louisville, Kentucky. The race track is one and one-quarter miles. The following table shows the winners since 1990, their margin of victory, the winning time, and the payoff on a $2 bet.

Year	Winner	Winning Margin (lengths)	Winning Time (minutes)	Payoff on a $2 Win Bet
1990	Unbridled	3.5	2.03333	10.80
1991	Strike the Gold	1.75	2.05000	4.80
1992	Lil E. Tee	1	2.05000	16.80
1993	Sea Hero	2.5	2.04000	12.90
1994	Go For Gin	2	2.06000	9.10
1995	Thunder Gulch	2.25	2.02000	24.50
1996	Grindstone	nose	2.01667	5.90
1997	Silver Charm	head	2.04000	4.00
1998	Real Quiet	0.5	2.03667	8.40
1999	Charismatic	neck	2.05333	31.30
2000	Fusaichi Pegasus	1.5	2.02000	2.30
2001	Monarchos	4.75	1.99950	10.50
2002	War Emblem	4	2.01883	20.50
2003	Funny Cide	1.75	2.01983	12.80
2004	Smarty Jones	2.75	2.06767	4.10
2005	Giacomo	0.5	2.04583	50.30
2006	Barbaro	6.5	2.02267	6.10
2007	Street Sense	2.25	2.03617	4.90
2008	Big Brown	4.75	2.03033	6.80
2009	Mine That Bird	6.75	2.04433	103.20
2010	Super Saver	2.50	2.07417	18.00
2011	Animal Kingdom	2.75	2.034	43.80
2012	I'll Have Another	1.5	2.03050	32.60
2013	Orb	2.5	2.04817	12.80
2014	California Chrome	1.75	2.0610	7.00
2015	American Pharaoh	1.00	2.05033	7.80

a. Determine the mean and median for the variables winning time and payoff on a $2 bet.
b. Determine the range and standard deviation of the variables winning time and payoff on a $2 bet.
c. Refer to the variable winning margin. What is the level of measurement? What measure of location would be most appropriate?

83. FILE The manager of the local Walmart Supercenter is studying the number of items purchased by customers in the evening hours. Listed below is the number of items for a sample of 30 customers.

15	8	6	9	9	4	18	10	10	12
12	4	7	8	12	10	10	11	9	13
5	6	11	14	5	6	6	5	13	5

a. Find the mean and the median of the number of items.
b. Find the range and the standard deviation of the number of items.
c. Organize the number of items into a frequency distribution. You may want to review the guidelines in Chapter 2 for establishing the class interval and the number of classes.
d. Find the mean and the standard deviation of the data organized into a frequency distribution. Compare these values with those computed in part (a). Why are they different?

84. The following frequency distribution reports the electricity cost for a sample of 50 two-bedroom apartments in Albuquerque, New Mexico, during the month of May last year.

Electricity Cost	Frequency
$ 80 up to $100	3
100 up to 120	8
120 up to 140	12
140 up to 160	16
160 up to 180	7
180 up to 200	4
Total	50

a. Estimate the mean cost.
b. Estimate the standard deviation.
c. Use the Empirical Rule to estimate the proportion of costs within two standard deviations of the mean. What are these limits?

85. Bidwell Electronics Inc. recently surveyed a sample of employees to determine how far they lived from corporate headquarters. The results are shown below. Compute the mean and the standard deviation.

Distance (miles)	Frequency	*M*
0 up to 5	4	2.5
5 up to 10	15	7.5
10 up to 15	27	12.5
15 up to 20	18	17.5
20 up to 25	6	22.5

DATA ANALYTICS

86. FILE Refer to the North Valley Real Estate data and prepare a report on the sales prices of the homes. Be sure to answer the following questions in your report.
a. Around what values of price do the data tend to cluster? What is the mean sales price? What is the median sales price? Is one measure more representative of the typical sales prices than the others?
b. What is the range of sales prices? What is the standard deviation? About 95% of the sales prices are between what two values? Is the standard deviation a useful statistic for describing the dispersion of sales price?
c. Repeat (a) and (b) using FICO score.

87. FILE Refer to the Baseball 2016 data, which report information on the 30 Major League Baseball teams for the 2016 season. Refer to the variable team salary.

a. Prepare a report on the team salaries. Be sure to answer the following questions in your report.

1. Around what values do the data tend to cluster? Specifically what is the mean team salary? What is the median team salary? Is one measure more representative of the typical team salary than the others?

2. What is the range of the team salaries? What is the standard deviation? About 95% of the salaries are between what two values?

b. Refer to the information on the average salary for each year. In 2000 the average player salary was $1.99 million. By 2016 the average player salary had increased to $4.40 million. What was the rate of increase over the period?

88. FILE Refer to the Lincolnville School District bus data. Prepare a report on the maintenance cost for last month. Be sure to answer the following questions in your report.

a. Around what values do the data tend to cluster? Specifically what was the mean maintenance cost last month? What is the median cost? Is one measure more representative of the typical cost than the others?

b. What is the range of maintenance costs? What is the standard deviation? About 95% of the maintenance costs are between what two values?

4 Describing Data:

DISPLAYING AND EXPLORING DATA

© Denis Vrublevski/Shutterstock.com

▲ **MCGIVERN JEWELERS** recently posted an advertisement on a social media site reporting the shape, size, price, and cut grade for 33 of its diamonds in stock. Develop a box plot of the variable price and comment on the result. (See Exercise 37 and LO4-4.)

LEARNING OBJECTIVES

When you have completed this chapter, you will be able to:

- **LO4-1** Construct and interpret a dot plot.
- **LO4-2** Construct and describe a stem-and-leaf display.
- **LO4-3** Identify and compute measures of position.
- **LO4-4** Construct and analyze a box plot.
- **LO4-5** Compute and interpret the coefficient of skewness.
- **LO4-6** Create and interpret a scatter diagram.
- **LO4-7** Develop and explain a contingency table.

INTRODUCTION

Chapter 2 began our study of descriptive statistics. In order to transform raw or ungrouped data into a meaningful form, we organize the data into a frequency distribution. We present the frequency distribution in graphic form as a histogram or a frequency polygon. This allows us to visualize where the data tend to cluster, the largest and the smallest values, and the general shape of the data.

In Chapter 3, we first computed several measures of location, such as the mean, median, and mode. These measures of location allow us to report a typical value in the set of observations. We also computed several measures of dispersion, such as the range, variance, and standard deviation. These measures of dispersion allow us to describe the variation or the spread in a set of observations.

We continue our study of descriptive statistics in this chapter. We study (1) dot plots, (2) stem-and-leaf displays, (3) percentiles, and (4) box plots. These charts and statistics give us additional insight into where the values are concentrated as well as the general shape of the data. Then we consider bivariate data. In bivariate data, we observe two variables for each individual or observation. Examples include the number of hours a student studied and the points earned on an examination; if a sampled product meets quality specifications and the shift on which it is manufactured; or the amount of electricity used in a month by a homeowner and the mean daily high temperature in the region for the month. These charts and graphs provide useful insights as we use business analytics to enhance our understanding of data.

LO4-1
Construct and interpret a dot plot.

DOT PLOTS

Recall for the Applewood Auto Group data, we summarized the profit earned on the 180 vehicles sold with a frequency distribution using eight classes. When we organized the data into the eight classes, we lost the exact value of the observations. A **dot plot,** on the other hand, groups the data as little as possible, and we do not lose the identity of an individual observation. To develop a dot plot, we display a dot for each observation along a horizontal number line indicating the possible values of the data. If there are identical observations or the observations are too close to be shown individually, the dots are "piled" on top of each other. This allows us to see the shape of the distribution, the value about which the data tend to cluster, and the largest and smallest observations. Dot plots are most useful for smaller data sets, whereas histograms tend to be most useful for large data sets. An example will show how to construct and interpret dot plots.

EXAMPLE

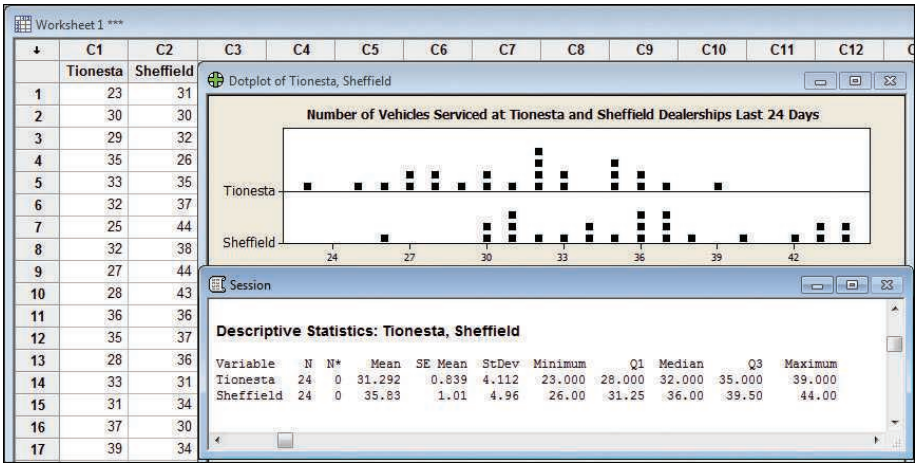
The service departments at Tionesta Ford Lincoln and Sheffield Motors Inc., two of the four Applewood Auto Group dealerships, were both open 24 days last month. Listed below is the number of vehicles serviced last month at the two dealerships. Construct dot plots and report summary statistics to compare the two dealerships.

Tionesta Ford Lincoln					
Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
23	33	27	28	39	26
30	32	28	33	35	32
29	25	36	31	32	27
35	32	35	37	36	30

Sheffield Motors Inc.					
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
31	35	44	36	34	37
30	37	43	31	40	31
32	44	36	34	43	36
26	38	37	30	42	33

SOLUTION

The Minitab system provides a dot plot and outputs the mean, median, maximum, and minimum values, and the standard deviation for the number of cars serviced at each dealership over the last 24 working days.



	C1	C2
	Tionesta	Sheffield
1	23	31
2	30	30
3	29	32
4	35	26
5	33	35
6	32	37
7	25	44
8	32	38
9	27	44
10	28	43
11	36	36
12	35	37
13	28	36
14	33	31
15	31	34
16	37	30
17	39	34

The dot plots, shown in the center of the output, graphically illustrate the distributions for each dealership. The plots show the difference in the location and dispersion of the observations. By looking at the dot plots, we can see that the number of vehicles serviced at the Sheffield dealership is more widely dispersed and has a larger mean than at the Tionesta dealership. Several other features of the number of vehicles serviced are:

- Tionesta serviced the fewest cars in any day, 23.
- Sheffield serviced 26 cars during their slowest day, which is 4 cars less than the next lowest day.
- Tionesta serviced exactly 32 cars on four different days.
- The numbers of cars serviced cluster around 36 for Sheffield and 32 for Tionesta.

From the descriptive statistics, we see Sheffield serviced a mean of 35.83 vehicles per day. Tionesta serviced a mean of 31.292 vehicles per day during the same period. So Sheffield typically services 4.54 more vehicles per day. There is also more dispersion, or variation, in the daily number of vehicles serviced at Sheffield than at Tionesta. How do we know this? The standard deviation is larger at Sheffield (4.96 vehicles per day) than at Tionesta (4.112 cars per day).

LO4-2
Construct and describe a stem-and-leaf display.

STEM-AND-LEAF DISPLAYS

In Chapter 2, we showed how to organize data into a frequency distribution so we could summarize the raw data into a meaningful form. The major advantage to organizing the data into a frequency distribution is we get a quick visual picture of the shape of the

distribution without doing any further calculation. To put it another way, we can see where the data are concentrated and also determine whether there are any extremely large or small values. There are two disadvantages, however, to organizing the data into a frequency distribution: (1) we lose the exact identity of each value and (2) we are not sure how the values within each class are distributed. To explain, the Theater of the Republic in Erie, Pennsylvania, books live theater and musical performances. The theater's capacity is 160 seats. Last year, among the forty-five performances, there were eight different plays and twelve different bands. The following frequency distribution shows that between eighty up to ninety people attended two of the forty-five performances; there were seven performances where ninety up to one hundred people attended. However, is the attendance within this class clustered about 90, spread evenly throughout the class, or clustered near 99? We cannot tell.

Attendance	Frequency
80 up to 90	2
90 up to 100	7
100 up to 110	6
110 up to 120	9
120 up to 130	8
130 up to 140	7
140 up to 150	3
150 up to 160	3
Total	45

One technique used to display quantitative information in a condensed form and provide more information than the frequency distribution is the **stem-and-leaf display.** An advantage of the stem-and-leaf display over a frequency distribution is we do not lose the identity of each observation. In the above example, we would not know the identity of the values in the 90 up to 100 class. To illustrate the construction of a stem-and-leaf display using the number people attending each performance, suppose the seven observations in the 90 up to 100 class are 96, 94, 93, 94, 95, 96, and 97. The **stem** value is the leading digit or digits, in this case 9. The **leaves** are the trailing digits. The stem is placed to the left of a vertical line and the leaf values to the right.

The values in the 90 up to 100 class would appear as follows:

9	6	4	3	4	5	6	7

It is also customary to sort the values within each stem from smallest to largest. Thus, the second row of the stem-and-leaf display would appear as follows:

9	3	4	4	5	6	6	7

With the stem-and-leaf display, we can quickly observe that 94 people attended two performances and the number attending ranged from 93 to 97. A stem-and-leaf display is similar to a frequency distribution with more information, that is, the identity of the observations is preserved.

STEM-AND-LEAF DISPLAY A statistical technique to present a set of data. Each numerical value is divided into two parts. The leading digit(s) becomes the stem and the trailing digit the leaf. The stems are located along the vertical axis, and the leaf values are stacked against each other along the horizontal axis.

The following example explains the details of developing a stem-and-leaf display.

EXAMPLE

Listed in Table 4–1 is the number of people attending each of the 45 performances at the Theater of the Republic last year. Organize the data into a stem-and-leaf display. Around what values does attendance tend to cluster? What is the smallest attendance? The largest attendance?

TABLE 4–1 Number of People Attending Each of the 45 Performances at the Theater of the Republic

96	93	88	117	127	95	113	96	108	94	148	156
139	142	94	107	125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139	134	119	97	89
118	136	125	143	120	103	113	124	138			

SOLUTION

From the data in Table 4–1, we note that the smallest attendance is 88. So we will make the first stem value 8. The largest attendance is 156, so we will have the stem values begin at 8 and continue to 15. The first number in Table 4–1 is 96, which has a stem value of 9 and a leaf value of 6. Moving across the top row, the second value is 93 and the third is 88. After the first 3 data values are considered, the chart is as follows.

Stem	Leaf
8	8
9	6 3
10	
11	
12	
13	
14	
15	

Organizing all the data, the stem-and-leaf chart looks as follows.

Stem	Leaf
8	8 9
9	6 3 5 6 4 4 7
10	8 7 3 4 6 3
11	7 3 2 7 2 1 9 8 3
12	7 5 7 0 5 5 0 4
13	9 5 2 9 4 6 8
14	8 2 3
15	6 5 5

The usual procedure is to sort the leaf values from the smallest to largest. The last line, the row referring to the values in the 150s, would appear as:

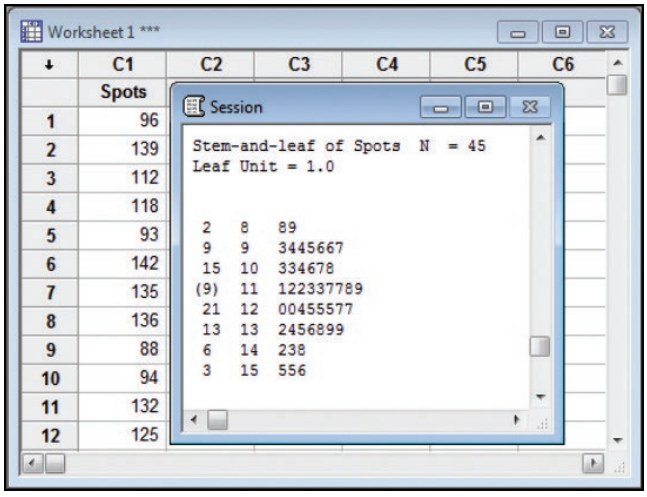
15	5	5	6

The final table would appear as follows, where we have sorted all of the leaf values.

Stem	Leaf
8	89
9	3445667
10	334678
11	122337789
12	00455577
13	2456899
14	238
15	556

You can draw several conclusions from the stem-and-leaf display. First, the minimum number of people attending is 88 and the maximum is 156. There were two performances with less than 90 people attending, and three performances with 150 or more. You can observe, for example, that for the three performances with more than 150 people attending, the actual attendances were 155, 155, and 156. The concentration of attendance is between 110 and 130. There were fifteen performances with attendance between 110 and 119 and eight performances between 120 and 129. We can also tell that within the 120 to 129 group the actual attendances were spread evenly throughout the class. That is, 120 people attended two performances, 124 people attended one performance, 125 people attended three performances, and 127 people attended two performances.

We also can generate this information on the Minitab software system. We have named the variable *Attendance*. The Minitab output is below. You can find the Minitab commands that will produce this output in Appendix C.



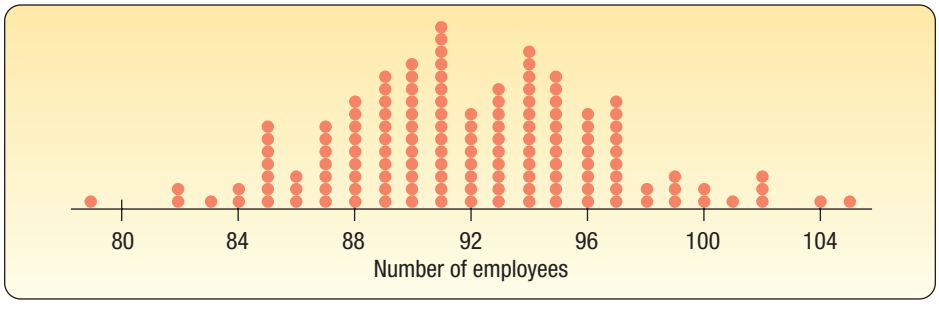
The Minitab solution provides some additional information regarding cumulative totals. In the column to the left of the stem values are numbers such as 2, 9, 15, and so on. The number 9 indicates there are 9 observations that have occurred before the value of 100. The number 15 indicates that 15 observations have occurred prior to 110. About halfway down the column the number 9 appears in parentheses. The parentheses indicate that the middle value or median appears in that row and there are nine values in this group. In this case, we describe the middle value as the value below which half of the observations occur. There are a total of 45 observations, so the middle value, if the data were arranged from smallest to largest, would be the 23rd observation; its value is 118. After the median, the values begin to decline. These values represent the "more than" cumulative totals. There are 21 observations of 120 or more, 13 of 130 or more, and so on.

Which is the better choice, a dot plot or a stem-and-leaf chart? This is really a matter of personal choice and convenience. For presenting data, especially with a large number of observations, you will find dot plots are more frequently used. You will see dot plots in analytical literature, marketing reports, and occasionally in annual reports. If you are doing a quick analysis for yourself, stem-and-leaf tallies are handy and easy, particularly on a smaller set of data.

SELF-REVIEW 4–1

© Somos/Veer/Getty Images RF

1. The number of employees at each of the 142 Home Depot stores in the Southeast region is shown in the following dot plot.



   (a) What are the maximum and minimum numbers of employees per store?
   (b) How many stores employ 91 people?
   (c) Around what values does the number of employees per store tend to cluster?

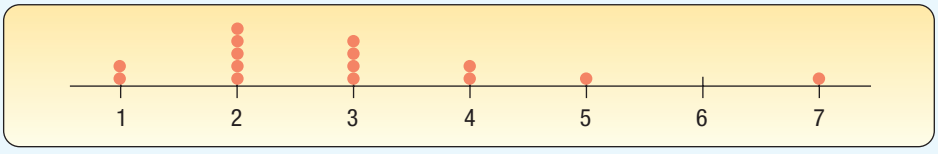
2. The rate of return for 21 stocks is:

8.3	9.6	9.5	9.1	8.8	11.2	7.7	10.1	9.9	10.8	
10.2	8.0	8.4	8.1	11.6	9.6	8.8	8.0	10.4	9.8	9.2

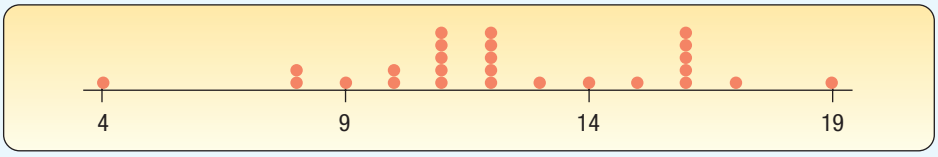
Organize this information into a stem-and-leaf display.
(a) How many rates are less than 9.0?
(b) List the rates in the 10.0 up to 11.0 category.
(c) What is the median?
(d) What are the maximum and the minimum rates of return?

EXERCISES

1. Describe the differences between a histogram and a dot plot. When might a dot plot be better than a histogram?
2. Describe the differences between a histogram and a stem-and-leaf display.
3. Consider the following chart.



   a. What is this chart called?
   b. How many observations are in the study?
   c. What are the maximum and the minimum values?
   d. Around what values do the observations tend to cluster?
4. The following chart reports the number of cell phones sold at a big-box retail store for the last 26 days.



   a. What are the maximum and the minimum numbers of cell phones sold in a day?
   b. What is a typical number of cell phones sold?
5. The first row of a stem-and-leaf chart appears as follows: 62 | 1 3 3 7 9. Assume whole number values.
   a. What is the "possible range" of the values in this row?
   b. How many data values are in this row?
   c. List the actual values in this row of data.
6. The third row of a stem-and-leaf chart appears as follows: 21 | 0 1 3 5 7 9. Assume whole number values.
   a. What is the "possible range" of the values in this row?
   b. How many data values are in this row?
   c. List the actual values in this row of data.
7. The following stem-and-leaf chart shows the number of units produced per day in a factory.

Stem	Leaf
3	8
4	
5	6
6	0133559
7	0236778
8	59
9	00156
10	36

   a. How many days were studied?
   b. How many observations are in the first class?

c. What are the minimum value and the maximum value?
d. List the actual values in the fourth row.
e. List the actual values in the second row.
f. How many values are less than 70?
g. How many values are 80 or more?
h. What is the median?
i. How many values are between 60 and 89, inclusive?

8. The following stem-and-leaf chart reports the number of prescriptions filled per day at the pharmacy on the corner of Fourth and Main Streets.

Stem	Leaf
12	689
13	123
14	6889
15	589
16	35
17	24568
18	268
19	13456
20	034679
21	2239
22	789
23	00179
24	8
25	13
26	
27	0

a. How many days were studied?
b. How many observations are in the last class?
c. What are the maximum and the minimum values in the entire set of data?
d. List the actual values in the fourth row.
e. List the actual values in the next to the last row.
f. On how many days were less than 160 prescriptions filled?
g. On how many days were 220 or more prescriptions filled?
h. What is the middle value?
i. How many days did the number of filled prescriptions range between 170 and 210?

9. FILE A survey of the number of phone calls made by a sample of 16 Verizon subscribers last week revealed the following information. Develop a stem-and-leaf chart. How many calls did a typical subscriber make? What were the maximum and the minimum number of calls made?

52	43	30	38	30	42	12	46	39
37	34	46	32	18	41	5		

10. FILE Aloha Banking Co. is studying ATM use in suburban Honolulu. Yesterday, for a sample of 30 ATM's, the bank counted the number of times each machine was used. The data is presented in the table. Develop a stem-and-leaf chart to summarize the data. What were the typical, minimum, and maximum number of times each ATM was used?

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

LO4-3
Identify and compute measures of position.

MEASURES OF POSITION

The standard deviation is the most widely used measure of dispersion. However, there are other ways of describing the variation or spread in a set of data. One method is to determine the *location* of values that divide a set of observations into equal parts. These measures include **quartiles, deciles,** and **percentiles.**

Quartiles divide a set of observations into four equal parts. To explain further, think of any set of values arranged from the minimum to the maximum. In Chapter 3, we called the middle value of a set of data arranged from the minimum to the maximum the median. That is, 50% of the observations are larger than the median and 50% are smaller. The median is a measure of location because it pinpoints the center of the data. In a similar fashion, **quartiles** divide a set of observations into four equal parts. The first quartile, usually labeled Q_1, is the value below which 25% of the observations occur, and the third quartile, usually labeled Q_3, is the value below which 75% of the observations occur.

Similarly, **deciles** divide a set of observations into 10 equal parts and **percentiles** into 100 equal parts. So if you found that your GPA was in the 8th decile at your university, you could conclude that 80% of the students had a GPA lower than yours and 20% had a higher GPA. If your GPA was in the 92nd percentile, then 92% of students had a GPA less than your GPA and only 8% of students had a GPA greater than your GPA. Percentile scores are frequently used to report results on such national standardized tests as the SAT, ACT, GMAT (used to judge entry into many master of business administration programs), and LSAT (used to judge entry into law school).

Quartiles, Deciles, and Percentiles

To formalize the computational procedure, let L_p refer to the location of a desired percentile. So if we want to find the 92nd percentile we would use L_{92}, and if we wanted the median, the 50th percentile, then L_{50}. For a number of observations, n, the location of the *Pth* percentile, can be found using the formula:

LOCATION OF A PERCENTILE

$$L_p = (n + 1)\frac{P}{100} \qquad [4\text{–}1]$$

An example will help to explain further.

EXAMPLE

Morgan Stanley is an investment company with offices located throughout the United States. Listed below are the commissions earned last month by a sample of 15 brokers at the Morgan Stanley office in Oakland, California.

$2,038	$1,758	$1,721	$1,637	$2,097	$2,047	$2,205	$1,787	$2,287
1,940	2,311	2,054	2,406	1,471	1,460			

Locate the median, the first quartile, and the third quartile for the commissions earned.

SOLUTION

The first step is to sort the data from the smallest commission to the largest.

$1,460	$1,471	$1,637	$1,721	$1,758	$1,787	$1,940	$2,038
2,047	2,054	2,097	2,205	2,287	2,311	2,406	

© Ramin Talaie/Getty Images

The median value is the observation in the center and is the same as the 50th percentile, so P equals 50. So the median or L_{50} is located at $(n + 1)(50/100)$, where n is the number of observations. In this case, that is position number 8, found by $(15 + 1)(50/100)$. The eighth-largest commission is $2,038. So we conclude this is the median and that half the brokers earned commissions more than $2,038 and half earned less than $2,038. The result using formula (4–1) to find the median is the same as the method presented in Chapter 3.

Recall the definition of a quartile. Quartiles divide a set of observations into four equal parts. Hence 25% of the observations will be less than the first quartile. Seventy-five percent of the observations will be less than the third quartile. To locate the first quartile, we use formula (4–1), where $n = 15$ and $P = 25$:

$$L_{25} = (n + 1)\frac{P}{100} = (15 + 1)\frac{25}{100} = 4$$

and to locate the third quartile, $n = 15$ and $P = 75$:

$$L_{75} = (n + 1)\frac{P}{100} = (15 + 1)\frac{75}{100} = 12$$

Therefore, the first and third quartile values are located at positions 4 and 12, respectively. The fourth value in the ordered array is $1,721 and the twelfth is $2,205. These are the first and third quartiles.

In the above example, the location formula yielded a whole number. That is, we wanted to find the first quartile and there were 15 observations, so the location formula indicated we should find the fourth ordered value. What if there were 20 observations in the sample, that is $n = 20$, and we wanted to locate the first quartile? From the location formula (4–1):

$$L_{25} = (n + 1)\frac{P}{100} = (20 + 1)\frac{25}{100} = 5.25$$

We would locate the fifth value in the ordered array and then move .25 of the distance between the fifth and sixth values and report that as the first quartile. Like the median, the quartile does not need to be one of the actual values in the data set.

To explain further, suppose a data set contained the six values 91, 75, 61, 101, 43, and 104. We want to locate the first quartile. We order the values from the minimum to the maximum: 43, 61, 75, 91, 101, and 104. The first quartile is located at

$$L_{25} = (n + 1)\frac{P}{100} = (6 + 1)\frac{25}{100} = 1.75$$

The position formula tells us that the first quartile is located between the first and the second values and it is .75 of the distance between the first and the second values. The first value is 43 and the second is 61. So the distance between these two values is 18. To locate the first quartile, we need to move .75 of the distance between the first and second values, so .75(18) = 13.5. To complete the procedure, we add 13.5 to the first value, 43, and report that the first quartile is 56.5.

We can extend the idea to include both deciles and percentiles. To locate the 23rd percentile in a sample of 80 observations, we would look for the 18.63 position.

$$L_{23} = (n + 1)\frac{P}{100} = (80 + 1)\frac{23}{100} = 18.63$$

STATISTICS IN ACTION

John W. Tukey (1915–2000) received a PhD in mathematics from Princeton in 1939. However, when he joined the Fire Control Research Office during World War II, his interest in abstract mathematics shifted to applied statistics. He developed effective numerical and graphical methods for studying patterns in data. Among the graphics he developed are the stem-and-leaf diagram and the box-and-whisker plot or box plot. From 1960 to 1980, Tukey headed the statistical division of NBC's election night vote projection team. He became renowned in 1960 for preventing an early call of victory for Richard Nixon in the presidential election won by John F. Kennedy.

To find the value corresponding to the 23rd percentile, we would locate the 18th value and the 19th value and determine the distance between the two values. Next, we would multiply this difference by 0.63 and add the result to the smaller value. The result would be the 23rd percentile.

Statistical software is very helpful when describing and summarizing data. Excel, Minitab, and MegaStat, a statistical analysis Excel add-in, all provide summary statistics that include quartiles. For example, the Minitab summary of the Morgan Stanley commission data, shown below, includes the first and third quartiles, and other statistics. Based on the reported quartiles, 25% of the commissions earned were less than $1,721 and 75% were less than $2,205. These are the same values we calculated using formula (4–1).

Sheet1 ***

	C1
	Commissions
1	1460
2	1471
3	1637
4	1721

Session

Descriptive Statistics: Commissions

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Commissions	15	0	1947.9	77.1	298.8	1460.0	1721.0	2038.0	2205.0	2406.0

Morgan Stanley Commissions
1460
2047
1471
2054
1637
2097
1721
2205
1758
2287
1787
2311
1940
2406
2038

Equation 4-1	
Quartile 1	1721
Quartile 3	2205

Alternate Method	
Quartile 1	1739.5
Quartile 3	2151

There are ways other than formula (4–1) to locate quartile values. For example, another method uses $0.25n + 0.75$ to locate the position of the first quartile and $0.75n + 0.25$ to locate the position of the third quartile. We will call this the *Excel Method*. In the Morgan Stanley data, this method would place the first quartile at position 4.5 (.25 × 15 + .75) and the third quartile at position 11.5 (.75 × 15 + .25). The first quartile would be interpolated as 0.5, or one-half the difference between the fourth- and the fifth-ranked values. Based on this method, the first quartile is $1739.5, found by ($1,721 + 0.5[$1,758 − $1,721]). The third quartile, at position 11.5, would be $2,151, or one-half the distance between the eleventh- and the twelfth-ranked values, found by ($2,097 + 0.5[$2,205 − $2,097]). Excel, as shown in the Morgan Stanley and Applewood examples, can compute quartiles using either of the two methods. **Please note the text uses formula (4–1) to calculate quartiles.**

Is the difference between the two methods important? No. Usually it is just a nuisance. In general, both methods calculate values that will support the statement that approximately 25% of the values are less than the value of the first quartile, and approximately 75% of the data values are less than the value of the third quartile. When the sample is large, the difference in the results from the two methods is small. For example, in the Applewood Auto Group data there are 180 vehicles. The quartiles computed using both methods are shown to the left. Based on the variable profit, 45 of the 180 values (25%) are less than both values of the first quartile, and 135 of the 180 values (75%) are less than both values of the third quartile.

Applewood	
Age	Profit
21	$1,387
23	$1,754
24	$1,817
25	$1,040
26	$1,273
27	$1,529
27	$3,082
28	$1,951
28	$2,692
29	$1,342

Equation 4-1	
Quartile 1	1415.5
Quartile 3	2275.5

Alternate Method	
Quartile 1	1422.5
Quartile 3	2268.5

When using Excel, be careful to understand the method used to

calculate quartiles. Excel 2013 and Excel 2016 offer both methods. The Excel function, ***Quartile.exc,*** will result in the same answer as Equation 4–1. The Excel function, *Quartile.inc*, will result in the Excel Method answers.

SELF-REVIEW 4–2

The Quality Control department of Plainsville Peanut Company is responsible for checking the weight of the 8-ounce jar of peanut butter. The weights of a sample of nine jars produced last hour are:

7.69	7.72	7.8	7.86	7.90	7.94	7.97	8.06	8.09

(a) What is the median weight?
(b) Determine the weights corresponding to the first and third quartiles.

EXERCISES

11. FILE Determine the median and the first and third quartiles in the following data.

46	47	49	49	51	53	54	54	55	55	59

12. FILE Determine the median and the first and third quartiles in the following data.

5.24	6.02	6.67	7.30	7.59	7.99	8.03	8.35	8.81	9.45
9.61	10.37	10.39	11.86	12.22	12.71	13.07	13.59	13.89	15.42

13. FILE The Thomas Supply Company Inc. is a distributor of gas-powered generators. As with any business, the length of time customers take to pay their invoices is important. Listed below, arranged from smallest to largest, is the time, in days, for a sample of The Thomas Supply Company Inc. invoices.

13	13	13	20	26	27	31	34	34	34	35	35	36	37	38
41	41	41	45	47	47	47	50	51	53	54	56	62	67	82

a. Determine the first and third quartiles.
b. Determine the second decile and the eighth decile.
c. Determine the 67th percentile.

14. FILE Kevin Horn is the national sales manager for National Textbooks Inc. He has a sales staff of 40 who visit college professors all over the United States. Each Saturday morning he requires his sales staff to send him a report. This report includes, among other things, the number of professors visited during the previous week. Listed below, ordered from smallest to largest, are the number of visits last week.

38	40	41	45	48	48	50	50	51	51	52	52	53	54	55	55	55	56	56	57
59	59	59	62	62	62	63	64	65	66	66	67	67	69	69	71	77	78	79	79

a. Determine the median number of calls.
b. Determine the first and third quartiles.
c. Determine the first decile and the ninth decile.
d. Determine the 33rd percentile.

LO4-4
Construct and analyze a box plot.

BOX PLOTS

A **box plot** is a graphical display, based on quartiles, that helps us picture a set of data. To construct a box plot, we need only five statistics: the minimum value, Q_1 (the first quartile), the median, Q_3 (the third quartile), and the maximum value. An example will help to explain.

EXAMPLE

Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

Minimum value = 13 minutes

Q_1 = 15 minutes

Median = 18 minutes

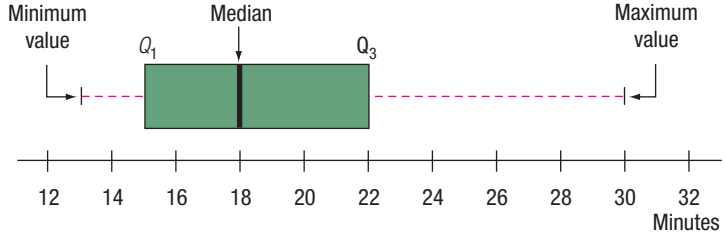
Q_3 = 22 minutes

Maximum value = 30 minutes

Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

SOLUTION

The first step in drawing a box plot is to create an appropriate scale along the horizontal axis. Next, we draw a box that starts at Q_1 (15 minutes) and ends at Q_3 (22 minutes). Inside the box we place a vertical line to represent the median (18 minutes). Finally, we extend horizontal lines from the box out to the minimum value (13 minutes) and the maximum value (30 minutes). These horizontal lines outside of the box are sometimes called "whiskers" because they look a bit like a cat's whiskers.



The box plot also shows the interquartile range of delivery times between Q1 and Q3. The **interquartile range** is 7 minutes and indicates that 50% of the deliveries are between 15 and 22 minutes.

The box plot also reveals that the distribution of delivery times is positively skewed. In Chapter 3, we defined skewness as the lack of symmetry in a set of data. How do we know this distribution is positively skewed? In this case, there are actually two pieces of information that suggest this. First, the dashed line to the right of the box from 22 minutes (Q_3) to the maximum time of 30 minutes is longer than the dashed line from the left of 15 minutes (Q_1) to the minimum value of 13 minutes. To put it another way,

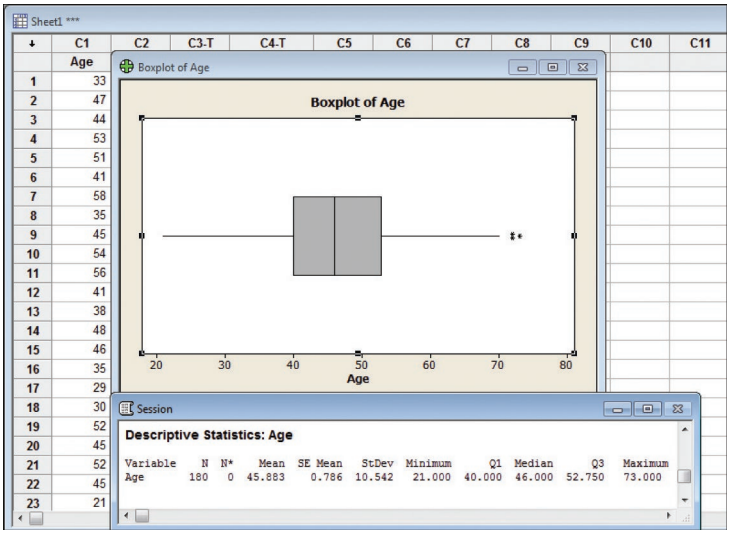
the 25% of the data larger than the third quartile is more spread out than the 25% less than the first quartile. A second indication of positive skewness is that the median is not in the center of the box. The distance from the first quartile to the median is smaller than the distance from the median to the third quartile. We know that the number of delivery times between 15 minutes and 18 minutes is the same as the number of delivery times between 18 minutes and 22 minutes.

EXAMPLE

Refer to the Applewood Auto Group data. Develop a box plot for the variable age of the buyer. What can we conclude about the distribution of the age of the buyer?

SOLUTION

Minitab was used to develop the following chart and summary statistics.



The median age of the purchaser is 46 years, 25% of the purchasers are less than 40 years of age, and 25% are more than 52.75 years of age. Based on the summary information and the box plot, we conclude:

- Fifty percent of the purchasers are between the ages of 40 and 52.75 years.
- The distribution of ages is fairly symmetric. There are two reasons for this conclusion. The length of the whisker above 52.75 years (Q_3) is about the same length as the whisker below 40 years (Q_1). Also, the area in the box between 40 years and the median of 46 years is about the same as the area between the median and 52.75.

There are three asterisks (*) above 70 years. What do they indicate? In a box plot, an asterisk identifies an **outlier.** An outlier is a value that is inconsistent with the rest of the data. It is defined as a value that is more than 1.5 times the interquartile range smaller than Q_1 or larger than Q_3. In this example, an outlier would be a value larger than 71.875 years, found by:

$$Outlier > Q_3 + 1.5(Q_3 - Q_1) = 52.75 + 1.5(52.75 - 40) = 71.875$$

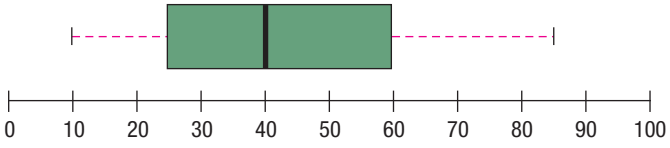
An outlier would also be a value less than 20.875 years.

$$Outlier < Q_1 - 1.5(Q_3 - Q_1) = 40 - 1.5(52.75 - 40) = 20.875$$

From the box plot, we conclude there are three purchasers 72 years of age or older and none less than 21 years of age. Technical note: In some cases, a single asterisk may represent more than one observation because of the limitations of the software and space available. It is a good idea to check the actual data. In this instance, there are three purchasers 72 years old or older; two are 72 and one is 73.

SELF-REVIEW 4–3

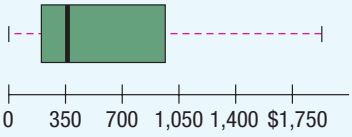
The following box plot shows the assets in millions of dollars for credit unions in Seattle, Washington.



What are the smallest and largest values, the first and third quartiles, and the median? Would you agree that the distribution is symmetrical? Are there any outliers?

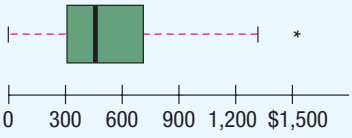
EXERCISES

15. The box plot below shows the amount spent for books and supplies per year by students at four-year public colleges.



a. Estimate the median amount spent.
b. Estimate the first and third quartiles for the amount spent.
c. Estimate the interquartile range for the amount spent.
d. Beyond what point is a value considered an outlier?
e. Identify any outliers and estimate their value.
f. Is the distribution symmetrical or positively or negatively skewed?

16. The box plot shows the undergraduate in-state tuition per credit hour at four-year public colleges.



a. Estimate the median.
b. Estimate the first and third quartiles.
c. Determine the interquartile range.
d. Beyond what point is a value considered an outlier?
e. Identify any outliers and estimate their value.
f. Is the distribution symmetrical or positively or negatively skewed?

17. In a study of the gasoline mileage of model year 2016 automobiles, the mean miles per gallon was 27.5 and the median was 26.8. The smallest value in the study was 12.70 miles per gallon, and the largest was 50.20. The first and third quartiles were 17.95 and 35.45 miles per gallon, respectively. Develop a box plot and comment on the distribution. Is it a symmetric distribution?

18. FILE A sample of 28 time shares in the Orlando, Florida, area revealed the following daily charges for a one-bedroom suite. For convenience, the data are ordered from smallest to largest. Construct a box plot to represent the data. Comment on the distribution. Be sure to identify the first and third quartiles and the median.

$116	$121	$157	$192	$207	$209	$209
229	232	236	236	239	243	246
260	264	276	281	283	289	296
307	309	312	317	324	341	353

LO4-5
Compute and interpret the coefficient of skewness.

SKEWNESS

In Chapter 3, we described measures of central location for a distribution of data by reporting the mean, median, and mode. We also described measures that show the amount of spread or variation in a distribution, such as the range and the standard deviation.

Another characteristic of a distribution is the shape. There are four shapes commonly observed: symmetric, positively skewed, negatively skewed, and bimodal. In a **symmetric** distribution the mean and median are equal and the data values are evenly spread around these values. The shape of the distribution below the mean and median is a mirror image of distribution above the mean and median. A distribution of values is **skewed to the right** or **positively skewed** if there is a single peak, but the values extend much farther to the right of the peak than to the left of the peak. In this case, the mean is larger than the median. In a **negatively skewed** distribution there is a single peak, but the observations extend farther to the left, in the negative direction, than to the right. In a negatively skewed distribution, the mean is smaller than the median. Positively skewed distributions are more common. Salaries often follow this pattern. Think of the salaries of those employed in a small company of about 100 people. The president and a few top executives would have very large salaries relative to the other workers and hence the distribution of salaries would exhibit positive skewness. A **bimodal distribution** will have two or more peaks. This is often the case when the values are from two or more populations. This information is summarized in Chart 4–1.

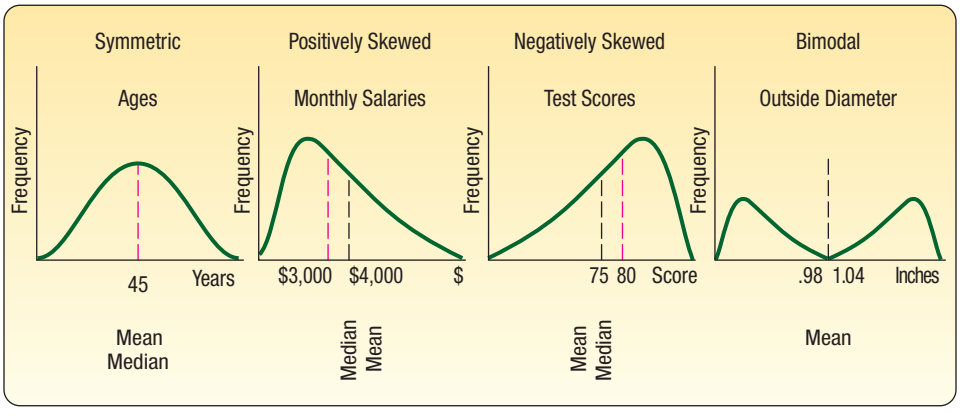


CHART 4–1 Shapes of Frequency Polygons

There are several formulas in the statistical literature used to calculate skewness. The simplest, developed by Professor Karl Pearson (1857–1936), is based on the difference between the mean and the median.

STATISTICS IN ACTION

The late Stephen Jay Gould (1941–2002) was a professor of zoology and professor of geology at Harvard University. In 1982, he was diagnosed with cancer and had an expected survival time of 8 months. However, never to be discouraged, his research showed that the distribution of survival time is dramatically skewed to the right and showed that not only do 50% of similar cancer patients survive more than 8 months, but that the survival time could be years rather than months! In fact, Dr. Gould lived another 20 years. Based on his experience, he wrote a widely published essay titled "The Median Is Not the Message."

PEARSON'S COEFFICIENT OF SKEWNESS

$$sk = \frac{3(\bar{x} - \text{Median})}{s} \qquad [4\text{–}2]$$

Using this relationship, the coefficient of skewness can range from −3 up to 3. A value near −3, such as −2.57, indicates considerable negative skewness. A value such as 1.63 indicates moderate positive skewness. A value of 0, which will occur when the mean and median are equal, indicates the distribution is symmetrical and there is no skewness present.

In this text, we present output from Minitab and Excel. Both of these software packages compute a value for the coefficient of skewness based on the cubed deviations from the mean. The formula is:

SOFTWARE COEFFICIENT OF SKEWNESS

$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{x - \bar{x}}{s}\right)^3\right] \qquad [4\text{–}3]$$

Formula (4–3) offers an insight into skewness. The right-hand side of the formula is the difference between each value and the mean, divided by the standard deviation. That is the portion $(x - \bar{x})/s$ of the formula. This idea is called **standardizing.** We will discuss the idea of standardizing a value in more detail in Chapter 7 when we describe the normal probability distribution. At this point, observe that the result is to report the difference between each value and the mean in units of the standard deviation. If this difference is positive, the particular value is larger than the mean; if the value is negative, the standardized quantity is smaller than the mean. When we cube these values, we retain the information on the direction of the difference. Recall that in the formula for the standard deviation [see formula (3–10)] we squared the difference between each value and the mean, so that the result was all nonnegative values.

If the set of data values under consideration is symmetric, when we cube the standardized values and sum over all the values, the result would be near zero. If there are several large values, clearly separate from the others, the sum of the cubed differences would be a large positive value. If there are several small values clearly separate from the others, the sum of the cubed differences will be negative.

An example will illustrate the idea of skewness.

EXAMPLE

Following are the earnings per share for a sample of 15 software companies for the year 2016. The earnings per share are arranged from smallest to largest.

$0.09	$0.13	$0.41	$0.51	$ 1.12	$ 1.20	$ 1.49	$3.18
3.50	6.36	7.83	8.92	10.13	12.99	16.40	

Compute the mean, median, and standard deviation. Find the coefficient of skewness using Pearson's estimate and the software methods. What is your conclusion regarding the shape of the distribution?

SOLUTION

These are sample data, so we use formula (3–2) to determine the mean

$$\bar{x} = \frac{\Sigma x}{n} = \frac{\$74.26}{15} = \$4.95$$

The median is the middle value in a set of data, arranged from smallest to largest. In this case, there is an odd-number of observations, so the middle value is the median. It is \$3.18.

We use formula (3–10) on page 78 to determine the sample standard deviation.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \cdots + (\$16.40 - \$4.95)^2}{15 - 1}} = \$5.22$$

Pearson's coefficient of skewness is 1.017, found by

$$sk = \frac{3(\bar{x} - \text{Median})}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

This indicates there is moderate positive skewness in the earnings per share data.

We obtain a similar, but not exactly the same, value from the software method. The details of the calculations are shown in Table 4–2. To begin, we find the difference between each earnings per share value and the mean and divide this result by the standard deviation. We have referred to this as standardizing. Next, we cube, that is, raise to the third power, the result of the first step. Finally, we sum the cubed values. The details for the first company, that is, the company with an earnings per share of \$0.09, are:

$$\left(\frac{x - \bar{x}}{s}\right)^3 = \left(\frac{0.09 - 4.95}{5.22}\right)^3 = (-0.9310)^3 = -0.8070$$

TABLE 4–2 Calculation of the Coefficient of Skewness

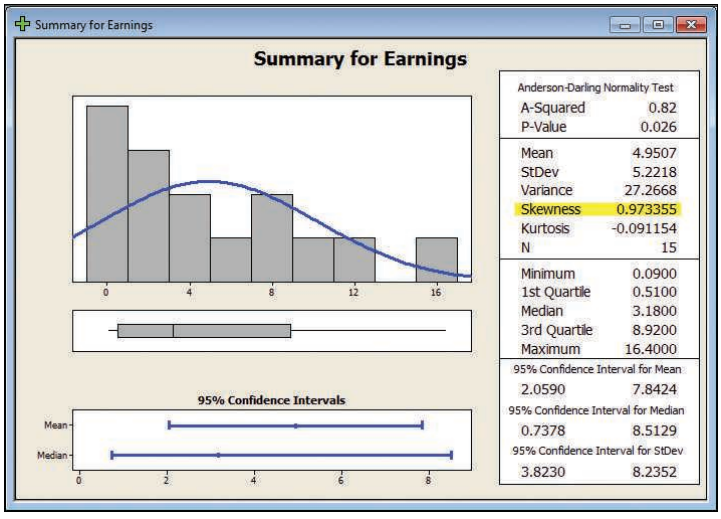
Earnings per Share	$\frac{(x - \bar{x})}{s}$	$\left(\frac{x - \bar{x}}{s}\right)^3$
0.09	−0.9310	−0.8070
0.13	−0.9234	−0.7873
0.41	−0.8697	−0.6579
0.51	−0.8506	−0.6154
1.12	−0.7337	−0.3950
1.20	−0.7184	−0.3708
1.49	−0.6628	−0.2912
3.18	−0.3391	−0.0390
3.50	−0.2778	−0.0214
6.36	0.2701	0.0197
7.83	0.5517	0.1679
8.92	0.7605	0.4399
10.13	0.9923	0.9772
12.99	1.5402	3.6539
16.40	2.1935	10.5537
		11.8274

When we sum the 15 cubed values, the result is 11.8274. That is, the term $\Sigma[(x - \bar{x})/s]^3 = 11.8274$. To find the coefficient of skewness, we use formula (4–3), with $n = 15$.

$$sk = \frac{n}{(n - 1)(n - 2)} \Sigma \left(\frac{x - \bar{x}}{s}\right)^3 = \frac{15}{(15 - 1)(15 - 2)} (11.8274) = 0.975$$

We conclude that the earnings per share values are somewhat positively skewed. The following Minitab summary reports the descriptive measures, such as

the mean, median, and standard deviation of the earnings per share data. Also included are the coefficient of skewness and a histogram with a bell-shaped curve superimposed.



SELF-REVIEW 4–4

A sample of five data entry clerks employed in the Horry County Tax Office revised the following number of tax records last hour: 73, 98, 60, 92, and 84.

(a) Find the mean, median, and the standard deviation.
(b) Compute the coefficient of skewness using Pearson's method.
(c) Calculate the coefficient of skewness using the software method.
(d) What is your conclusion regarding the skewness of the data?

EXERCISES

For Exercises 19–22:

a. Determine the mean, median, and the standard deviation.
b. Determine the coefficient of skewness using Pearson's method.
c. Determine the coefficient of skewness using the software method.

19. FILE The following values are the starting salaries, in $000, for a sample of five accounting graduates who accepted positions in public accounting last year.

36.0	26.0	33.0	28.0	31.0

20. FILE Listed below are the salaries, in $000, for a sample of 15 chief financial officers in the electronics industry.

$516.0	$548.0	$566.0	$534.0	$586.0	$529.0
546.0	523.0	538.0	523.0	551.0	552.0
486.0	558.0	574.0			

21. FILE Listed below are the commissions earned ($000) last year by the 15 sales representatives at Furniture Patch Inc.

$ 3.9	$ 5.7	$ 7.3	$10.6	$13.0	$13.6	$15.1	$15.8	$17.1
17.4	17.6	22.3	38.6	43.2	87.7			

22. FILE Listed below are the salaries for the 2016 New York Yankees Major League Baseball team.

Player	Salary	Player	Salary
CC Sabathia	$25,000,000	Dustin Ackley	$3,200,000
Mark Teixeira	23,125,000	Martin Prado	3,000,000
Masahiro Tanaka	22,000,000	Didi Gregorius	2,425,000
Jacoby Ellsbury	21,142,857	Aaron Hicks	574,000
Alex Rodriguez	21,000,000	Austin Romine	556,000
Brian McCann	17,000,000	Chasen Shreve	533,400
Carlos Beltran	15,000,000	Greg Bird	525,300
Brett Gardner	13,500,000	Luis Severino	521,300
Chase Headley	13,000,000	Bryan Mitchell	516,650
Aroldis Chapman	11,325,000	Kirby Yates	511,900
Andrew Miller	9,000,000	Mason Williams	509,700
Starlin Castro	7,857,143	Ronald Torreyes	508,600
Nathan Eovaldi	5,600,000	John Barbato	507,500
Michael Pineda	4,300,000	Dellin Betances	507,500
Ivan Nova	4,100,000	Luis Cessa	507,500

LO4-6
Create and interpret a scatter diagram.

DESCRIBING THE RELATIONSHIP BETWEEN TWO VARIABLES

In Chapter 2 and the first section of this chapter, we presented graphical techniques to summarize the distribution of a single variable. We used a histogram in Chapter 2 to summarize the profit on vehicles sold by the Applewood Auto Group. Earlier in this chapter, we used dot plots and stem-and-leaf displays to visually summarize a set of data. Because we are studying a single variable, we refer to this as **univariate** data.

© Steve Mason/Getty Images RF

There are situations where we wish to study and visually portray the relationship between two variables. When we study the relationship between two variables, we refer to the data as **bivariate.** Data analysts frequently wish to understand the relationship between two variables. Here are some examples:

- Tybo and Associates is a law firm that advertises extensively on local TV. The partners are considering increasing their advertising budget. Before doing so, they would like to know the relationship between the amount spent per month on advertising and the total amount of billings for that month. To put it another way, will increasing the amount spent on advertising result in an increase in billings?

- Coastal Realty is studying the selling prices of homes. What variables seem to be related to the selling price of homes? For example, do larger homes sell for more than smaller ones? Probably. So Coastal might study the relationship between the area in square feet and the selling price.
- Dr. Stephen Givens is an expert in human development. He is studying the relationship between the height of fathers and the height of their sons. That is, do tall fathers tend to have tall children? Would you expect LeBron James, the 6′8″, 250 pound professional basketball player, to have relatively tall sons?

One graphical technique we use to show the relationship between variables is called a **scatter diagram.**

To draw a scatter diagram, we need two variables. We scale one variable along the horizontal axis (*X*-axis) of a graph and the other variable along the vertical axis (*Y*-axis). Usually one variable depends to some degree on the other. In the third example above, the height of the son *depends* on the height of the father. So we scale the height of the father on the horizontal axis and that of the son on the vertical axis.

We can use statistical software, such as Excel, to perform the plotting function for us. *Caution:* You should always be careful of the scale. By changing the scale of either the vertical or the horizontal axis, you can affect the apparent visual strength of the relationship.

Following are three scatter diagrams (Chart 4–2). The one on the left shows a rather strong positive relationship between the age in years and the maintenance cost last year for a sample of 10 buses owned by the city of Cleveland, Ohio. Note that as the age of the bus increases, the yearly maintenance cost also increases. The example in the center, for a sample of 20 vehicles, shows a rather strong indirect relationship between the odometer reading and the auction price. That is, as the number of miles driven increases, the auction price decreases. The example on the right depicts the relationship between the height and yearly salary for a sample of 15 shift supervisors. This graph indicates there is little relationship between their height and yearly salary.

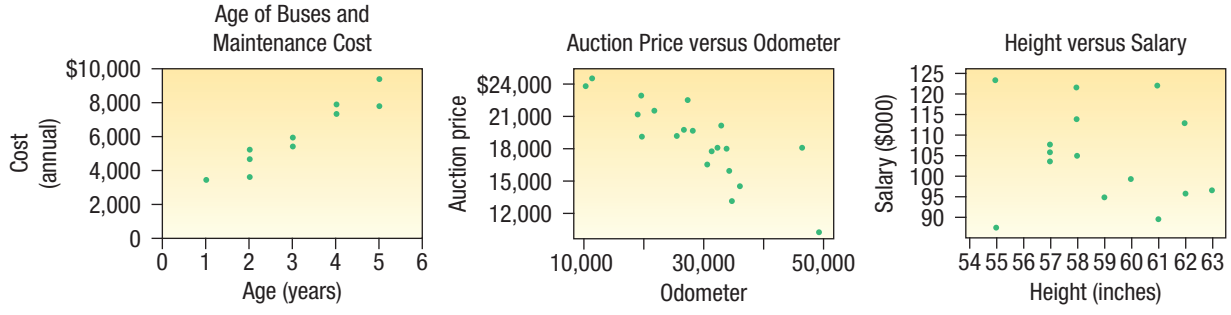


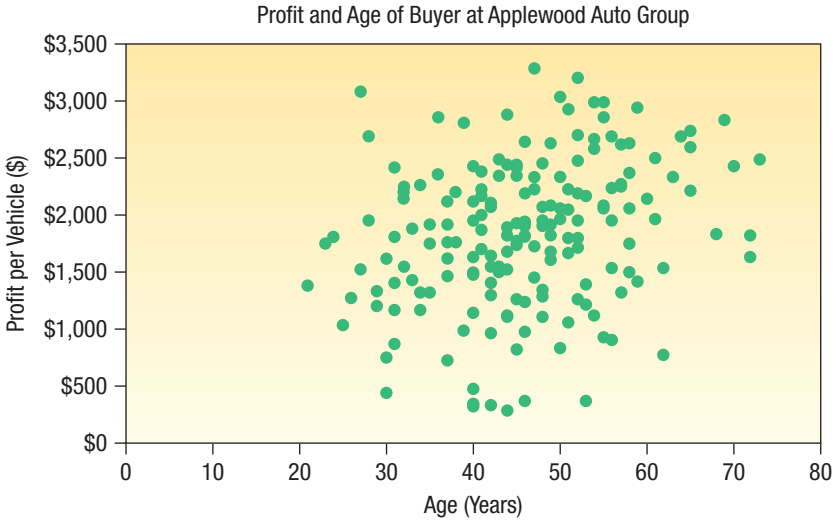
CHART 4–2 Three Examples of Scatter Diagrams.

EXAMPLE

In the introduction to Chapter 2, we presented data from the Applewood Auto Group. We gathered information concerning several variables, including the profit earned from the sale of 180 vehicles sold last month. In addition to the amount of profit on each sale, one of the other variables is the age of the purchaser. Is there a relationship between the profit earned on a vehicle sale and the age of the purchaser? Would it be reasonable to conclude that more profit is made on vehicles purchased by older buyers?

SOLUTION

We can investigate the relationship between vehicle profit and the age of the buyer with a scatter diagram. We scale age on the horizontal, or *X*-axis, and the profit on the vertical, or *Y*-axis. We assume profit depends on the age of the purchaser. As people age, they earn more income and purchase more expensive cars which, in turn, produce higher profits. We use Excel to develop the scatter diagram. The Excel commands are in Appendix C.



The scatter diagram shows a rather weak positive relationship between the two variables. It does not appear there is much relationship between the vehicle profit and the age of the buyer. In Chapter 13, we will study the relationship between variables more extensively, even calculating several numerical measures to express the relationship between variables.

In the preceding example, there is a weak positive, or direct, relationship between the variables. There are, however, many instances where there is a relationship between the variables, but that relationship is inverse or negative. For example:

- The value of a vehicle and the number of miles driven. As the number of miles increases, the value of the vehicle decreases.
- The premium for auto insurance and the age of the driver. Auto rates tend to be the highest for younger drivers and less for older drivers.
- For many law enforcement personnel, as the number of years on the job increases, the number of traffic citations decreases. This may be because personnel become more liberal in their interpretations or they may be in supervisor positions and not in a position to issue as many citations. But in any event, as age increases, the number of citations decreases.

LO4-7
Develop and explain a contingency table.

CONTINGENCY TABLES

A scatter diagram requires that both of the variables be at least interval scale. In the Applewood Auto Group example, both age and vehicle profit are ratio scale variables. Height is also ratio scale as used in the discussion of the relationship between the height of fathers and the height of their sons. What if we wish to study the relationship between two variables when one or both are nominal or ordinal scale? In this case, we tally the results in a **contingency table.**

CONTINGENCY TABLE A table used to classify observations according to two identifiable characteristics.

A contingency table is a cross-tabulation that simultaneously summarizes two variables of interest. For example:

- Students at a university are classified by gender and class (freshman, sophomore, junior, or senior).
- A product is classified as acceptable or unacceptable and by the shift (day, afternoon, or night) on which it is manufactured.
- A voter in a school bond referendum is classified as to party affiliation (Democrat, Republican, other) and the number of children that voter has attending school in the district (0, 1, 2, etc.).

EXAMPLE

There are four dealerships in the Applewood Auto Group. Suppose we want to compare the profit earned on each vehicle sold by the particular dealership. To put it another way, is there a relationship between the amount of profit earned and the dealership?

SOLUTION

In a contingency table, both variables only need to be nominal or ordinal. In this example, the variable dealership is a nominal variable and the variable profit is a ratio variable. To convert profit to an ordinal variable, we classify the variable profit into two categories, those cases where the profit earned is more than the median and those cases where it is less. On page 64, we calculated the median profit for all sales last month at Applewood Auto Group to be $1,882.50.

Contingency Table Showing the Relationship between Profit and Dealership

Above/Below Median Profit	Kane	Olean	Sheffield	Tionesta	Total
Above	25	20	19	26	90
Below	27	20	26	17	90
Total	52	40	45	43	180

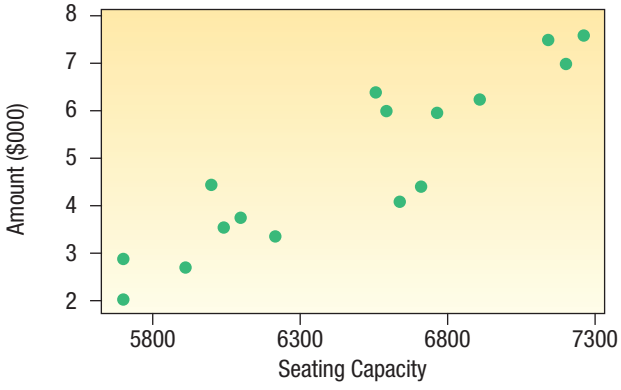
By organizing the information into a contingency table, we can compare the profit at the four dealerships. We observe the following:

- From the Total column on the right, 90 of the 180 cars sold had a profit above the median and half below. From the definition of the median, this is expected.
- For the Kane dealership, 25 out of the 52, or 48%, of the cars sold were sold for a profit more than the median.
- The percentage of profits above the median for the other dealerships are 50% for Olean, 42% for Sheffield, and 60% for Tionesta.

We will return to the study of contingency tables in Chapter 5 during the study of probability and in Chapter 15 during the study of nonparametric methods of analysis.

SELF-REVIEW 4–5

The rock group Blue String Beans is touring the United States. The following chart shows the relationship between concert seating capacity and revenue in $000 for a sample of concerts.



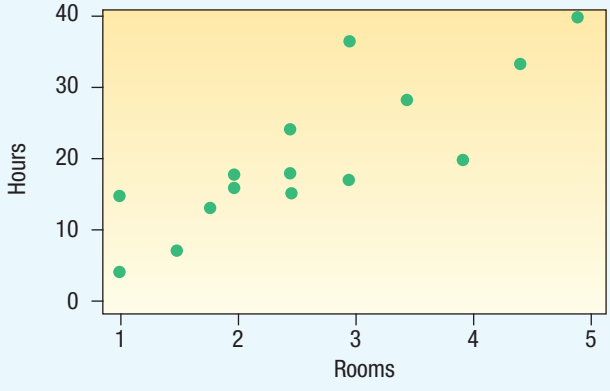
(a) What is the diagram called?
(b) How many concerts were studied?
(c) Estimate the revenue for the concert with the largest seating capacity.
(d) How would you characterize the relationship between revenue and seating capacity? Is it strong or weak, direct or inverse?

EXERCISES

23. FILE Develop a scatter diagram for the following sample data. How would you describe the relationship between the values?

x-Value	*y*-Value	*x*-Value	*y*-Value
10	6	11	6
8	2	10	5
9	6	7	2
11	5	7	3
13	7	11	7

24. Silver Springs Moving and Storage Inc. is studying the relationship between the number of rooms in a move and the number of labor hours required for the move. As part of the analysis, the CFO of Silver Springs developed the following scatter diagram.



a. How many moves are in the sample?
b. Does it appear that more labor hours are required as the number of rooms increases, or do labor hours decrease as the number of rooms increases?

25. The Director of Planning for Devine Dining Inc. wishes to study the relationship between the gender of a guest and whether the guest orders dessert. To investigate the relationship, the manager collected the following information on 200 recent customers.

	Gender		
Dessert Ordered	**Male**	**Female**	**Total**
Yes	32	15	47
No	68	85	153
Total	100	100	200

a. What is the level of measurement of the two variables?
b. What is the above table called?
c. Does the evidence in the table suggest men are more likely to order dessert than women? Explain why.

26. Ski Resorts of Vermont Inc. is considering a merger with Gulf Shores Beach Resorts Inc. of Alabama. The board of directors surveyed 50 stockholders concerning their position on the merger. The results are reported below.

	Opinion			
Number of Shares Held	**Favor**	**Oppose**	**Undecided**	**Total**
Under 200	8	6	2	16
200 up to 1,000	6	8	1	15
Over 1,000	6	12	1	19
Total	20	26	4	50

a. What level of measurement is used in this table?
b. What is this table called?
c. What group seems most strongly opposed to the merger?

CHAPTER SUMMARY

I. A dot plot shows the range of values on the horizontal axis and the number of observations for each value on the vertical axis.
  A. Dot plots report the details of each observation.
  B. They are useful for comparing two or more data sets.
II. A stem-and-leaf display is an alternative to a histogram.
  A. The leading digit is the stem and the trailing digit the leaf.
  B. The advantages of a stem-and-leaf display over a histogram include:
    1. The identity of each observation is not lost.
    2. The digits themselves give a picture of the distribution.
    3. The cumulative frequencies are also shown.
III. Measures of location also describe the shape of a set of observations.
  A. Quartiles divide a set of observations into four equal parts.
    1. Twenty-five percent of the observations are less than the first quartile, 50% are less than the second quartile, and 75% are less than the third quartile.
    2. The interquartile range is the difference between the third quartile and the first quartile.
  B. Deciles divide a set of observations into 10 equal parts and percentiles into 100 equal parts.

IV. A box plot is a graphic display of a set of data.
- **A.** A box is drawn enclosing the regions between the first quartile and the third quartile.
  - **1.** A line is drawn inside the box at the median value.
  - **2.** Dotted line segments are drawn from the third quartile to the largest value to show the highest 25% of the values and from the first quartile to the smallest value to show the lowest 25% of the values.
- **B.** A box plot is based on five statistics: the maximum and minimum values, the first and third quartiles, and the median.

V. The coefficient of skewness is a measure of the symmetry of a distribution.
- **A.** There are two formulas for the coefficient of skewness.
  - **1.** The formula developed by Pearson is:

$$sk = \frac{3(\bar{x} - \text{Median})}{s} \qquad \textbf{[4–2]}$$

  - **2.** The coefficient of skewness computed by statistical software is:

$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{x-\bar{x}}{s}\right)^3\right] \qquad \textbf{[4–3]}$$

VI. A scatter diagram is a graphic tool to portray the relationship between two variables.
- **A.** Both variables are measured with interval or ratio scales.
- **B.** If the scatter of points moves from the lower left to the upper right, the variables under consideration are directly or positively related.
- **C.** If the scatter of points moves from the upper left to the lower right, the variables are inversely or negatively related.

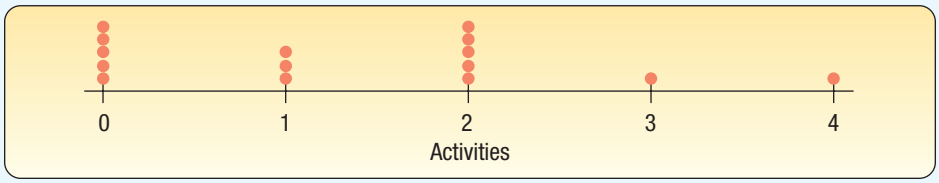
VII. A contingency table is used to classify nominal-scale observations according to two characteristics.

PRONUNCIATION KEY

SYMBOL	MEANING	PRONUNCIATION
L_p	Location of percentile	*L* sub *p*
Q_1	First quartile	*Q* sub 1
Q_3	Third quartile	*Q* sub 3

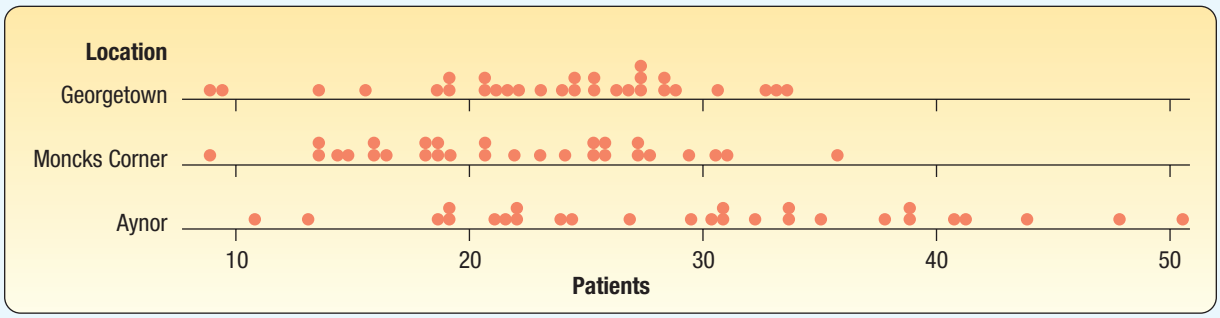
CHAPTER EXERCISES

27. A sample of students attending Southeast Florida University is asked the number of social activities in which they participated last week. The chart below was prepared from the sample data.



- **a.** What is the name given to this chart?
- **b.** How many students were in the study?
- **c.** How many students reported attending no social activities?

28. Doctor's Care is a walk-in clinic, with locations in Georgetown, Moncks Corner, and Aynor, at which patients may receive treatment for minor injuries, colds, and flu, as well as physical examinations. The following charts report the number of patients treated in each of the three locations last month.



Describe the number of patients served at the three locations each day. What are the maximum and minimum numbers of patients served at each of the locations?

29. FILE Below is the number of customers who visited Smith's True-Value hardware store in Bellville, Ohio, over the last twenty-three days. Make a stem-and-leaf display of this variable.

46	52	46	40	42	46	40	37	46	40	52	32	37	32	52
40	32	52	40	52	46	46	52							

30. FILE The top 25 companies (by market capitalization) operating in the Washington, DC, area along with the year they were founded and the number of employees are given below. Make a stem-and-leaf display of each of these variables and write a short description of your findings.

Company Name	Year Founded	Employees
AES Corp.	1981	30,000
American Capital Ltd.	1986	484
AvalonBay Communities Inc.	1978	1,767
Capital One Financial Corp.	1995	31,800
Constellation Energy Group Inc.	1816	9,736
Coventry Health Care Inc.	1986	10,250
Danaher Corp.	1984	45,000
Dominion Resources Inc.	1909	17,500
Fannie Mae	1938	6,450
Freddie Mac	1970	5,533
Gannett Co.	1906	49,675
General Dynamics Corp.	1952	81,000
Genworth Financial Inc.	2004	7,200
Harman International Industries Inc.	1980	11,246
Host Hotels & Resorts Inc.	1927	229
Legg Mason	1899	3,800
Lockheed Martin Corp.	1995	140,000
Marriott International Inc.	1927	151,000
MedImmune LLC	1988	2,516
NII Holdings Inc.	1996	7,748
Norfolk Southern Corp.	1982	30,594
Pepco Holdings Inc.	1896	5,057
Sallie Mae	1972	11,456
T. Rowe Price Group Inc.	1937	4,605
The Washington Post Co.	1877	17,100

31. FILE In recent years, due to low interest rates, many homeowners refinanced their home mortgages. Linda Lahey is a mortgage officer at Down River Federal Savings

and Loan. Below is the amount refinanced for 20 loans she processed last week. The data are reported in thousands of dollars and arranged from smallest to largest.

59.2	59.5	61.6	65.5	66.6	72.9	74.8	77.3	79.2
83.7	85.6	85.8	86.6	87.0	87.1	90.2	93.3	98.6
100.2	100.7							

a. Find the median, first quartile, and third quartile.
b. Find the 26th and 83rd percentiles.
c. Draw a box plot of the data.

32. FILE A study is made by the recording industry in the United States of the number of music CDs owned by 25 senior citizens and 30 young adults. The information is reported below.

Seniors									
28	35	41	48	52	81	97	98	98	99
118	132	133	140	145	147	153	158	162	174
177	180	180	187	188					

Young Adults									
81	107	113	147	147	175	183	192	202	209
233	251	254	266	283	284	284	316	372	401
417	423	490	500	507	518	550	557	590	594

a. Find the median and the first and third quartiles for the number of CDs owned by senior citizens. Develop a box plot for the information.
b. Find the median and the first and third quartiles for the number of CDs owned by young adults. Develop a box plot for the information.
c. Compare the number of CDs owned by the two groups.

33. FILE The corporate headquarters of *Bank.com,* an on-line banking company, is located in downtown Philadelphia. The director of human resources is making a study of the time it takes employees to get to work. The city is planning to offer incentives to each downtown employer if they will encourage their employees to use public transportation. Below is a listing of the time to get to work this morning according to whether the employee used public transportation or drove a car.

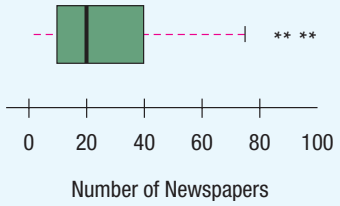
Public Transportation									
23	25	25	30	31	31	32	33	35	36
37	42								

Private									
32	32	33	34	37	37	38	38	38	39
40	44								

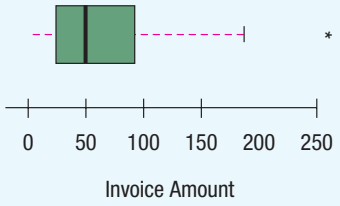
a. Find the median and the first and third quartiles for the time it took employees using public transportation. Develop a box plot for the information.
b. Find the median and the first and third quartiles for the time it took employees who drove their own vehicle. Develop a box plot for the information.
c. Compare the times of the two groups.

34. The following box plot shows the number of daily newspapers published in each state and the District of Columbia. Write a brief report summarizing the number published. Be sure to include information on the values of the first and third quartiles,

the median, and whether there is any skewness. If there are any outliers, estimate their value.



35. Walter Gogel Company is an industrial supplier of fasteners, tools, and springs. The amounts of its invoices vary widely, from less than $20.00 to more than $400.00. During the month of January the company sent out 80 invoices. Here is a box plot of these invoices. Write a brief report summarizing the invoice amounts. Be sure to include information on the values of the first and third quartiles, the median, and whether there is any skewness. If there are any outliers, approximate the value of these invoices.



36. FILE The American Society of PeriAnesthesia Nurses (ASPAN; www.aspan.org) is a national organization serving nurses practicing in ambulatory surgery, preanesthesia, and postanesthesia care. The organization consists of the 40 components listed below.

State/Region	Membership	State/Region	Membership
Alabama	95	New Jersey, Bermuda	517
Arizona	399	Alaska, Idaho, Montana, Oregon, Washington	708
Maryland, Delaware, DC	531	New York	891
Connecticut	239	Ohio	708
Florida	631	Oklahoma	171
Georgia	384	Arkansas	68
Hawaii	73	California	1,165
Illinois	562	New Mexico	79
Indiana	270	Pennsylvania	575
Iowa	117	Rhode Island	53
Kentucky	197	Colorado	409
Louisiana	258	South Carolina	237
Michigan	411	Texas	1,026
Massachusetts	480	Tennessee	167
Maine	97	Utah	67
Minnesota, Dakotas	289	Virginia	414
Missouri, Kansas	282	Vermont, New Hampshire	144
Mississippi	90	Wisconsin	311
Nebraska	115	West Virginia	62
North Carolina	542		
Nevada	106		

Use statistical software to answer the following questions.

a. Find the mean, median, and standard deviation of the number of members per component.

b. Find the coefficient of skewness, using the software. What do you conclude about the shape of the distribution of component size?
c. Compute the first and third quartiles using formula (4–1).
d. Develop a box plot. Are there any outliers? Which components are outliers? What are the limits for outliers?

37. FILE McGivern Jewelers is located in the Levis Square Mall just south of Toledo, Ohio. Recently it posted an advertisement on a social media site reporting the shape, size, price, and cut grade for 33 of its diamonds currently in stock. The information is reported below.

Shape	Size (carats)	Price	Cut Grade	Shape	Size (carats)	Price	Cut Grade
Princess	5.03	$44,312	Ideal cut	Round	0.77	$2,828	Ultra ideal cut
Round	2.35	20,413	Premium cut	Oval	0.76	3,808	Premium cut
Round	2.03	13,080	Ideal cut	Princess	0.71	2,327	Premium cut
Round	1.56	13,925	Ideal cut	Marquise	0.71	2,732	Good cut
Round	1.21	7,382	Ultra ideal cut	Round	0.70	1,915	Premium cut
Round	1.21	5,154	Average cut	Round	0.66	1,885	Premium cut
Round	1.19	5,339	Premium cut	Round	0.62	1,397	Good cut
Emerald	1.16	5,161	Ideal cut	Round	0.52	2,555	Premium cut
Round	1.08	8,775	Ultra ideal cut	Princess	0.51	1,337	Ideal cut
Round	1.02	4,282	Premium cut	Round	0.51	1,558	Premium cut
Round	1.02	6,943	Ideal cut	Round	0.45	1,191	Premium cut
Marquise	1.01	7,038	Good cut	Princess	0.44	1,319	Average cut
Princess	1.00	4,868	Premium cut	Marquise	0.44	1,319	Premium cut
Round	0.91	5,106	Premium cut	Round	0.40	1,133	Premium cut
Round	0.90	3,921	Good cut	Round	0.35	1,354	Good cut
Round	0.90	3,733	Premium cut	Round	0.32	896	Premium cut
Round	0.84	2,621	Premium cut				

a. Develop a box plot of the variable price and comment on the result. Are there any outliers? What is the median price? What are the values of the first and the third quartiles?
b. Develop a box plot of the variable size and comment on the result. Are there any outliers? What is the median price? What are the values of the first and the third quartiles?
c. Develop a scatter diagram between the variables price and size. Be sure to put price on the vertical axis and size on the horizontal axis. Does there seem to be an association between the two variables? Is the association direct or indirect? Does any point seem to be different from the others?
d. Develop a contingency table for the variables shape and cut grade. What is the most common cut grade? What is the most common shape? What is the most common combination of cut grade and shape?

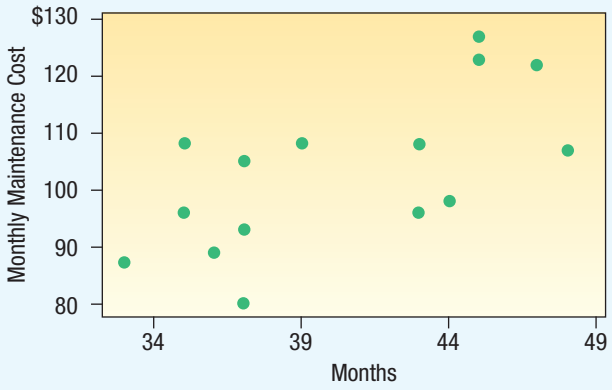
38. FILE Listed below is the amount of commissions earned last month for the eight members of the sales staff at Best Electronics. Calculate the coefficient of skewness using both methods. Hint: Use of a spreadsheet will expedite the calculations.

980.9	1,036.5	1,099.5	1,153.9	1,409.0	1,456.4	1,718.4	1,721.2

39. FILE Listed below is the number of car thefts in a large city over the last week. Calculate the coefficient of skewness using both methods. Hint: Use of a spreadsheet will expedite the calculations.

3	12	13	7	8	3	8

40. The manager of Information Services at Wilkin Investigations, a private investigation firm, is studying the relationship between the age (in months) of a combination printer, copier, and fax machine and its monthly maintenance cost. For a sample of 15 machines, the manager developed the following chart. What can the manager conclude about the relationship between the variables?



41. FILE An auto insurance company reported the following information regarding the age of a driver and the number of accidents reported last year. Develop a scatter diagram for the data and write a brief summary.

Age	Accidents	Age	Accidents
16	4	23	0
24	2	27	1
18	5	32	1
17	4	22	3

42. Wendy's offers eight different condiments (mustard, catsup, onion, mayonnaise, pickle, lettuce, tomato, and relish) on hamburgers. A store manager collected the following information on the number of condiments ordered and the age group of the customer. What can you conclude regarding the information? Who tends to order the most or least number of condiments?

	Age			
Number of Condiments	Under 18	18 up to 40	40 up to 60	60 or older
0	12	18	24	52
1	21	76	50	30
2	39	52	40	12
3 or more	71	87	47	28

43. Here is a table showing the number of employed and unemployed workers 20 years or older by gender in the United States.

	Number of Workers (000)	
Gender	Employed	Unemployed
Men	70,415	4,209
Women	61,402	3,314

a. How many workers were studied?
b. What percent of the workers were unemployed?
c. Compare the percent unemployed for the men and the women.

DATA ANALYTICS

44. FILE Refer to the North Valley real estate data recorded on homes sold during the last year. Prepare a report on the selling prices of the homes based on the answers to the following questions.
- **a.** Compute the minimum, maximum, median, and the first and the third quartiles of price. Create a box plot. Comment on the distribution of home prices.
- **b.** Develop a scatter diagram with price on the vertical axis and the size of the home on the horizontal. Is there a relationship between these variables? Is the relationship direct or indirect?
- **c.** For homes without a pool, develop a scatter diagram with price on the vertical axis and the size of the home on the horizontal. Do the same for homes with a pool. How do the relationships between price and size for homes without a pool and homes with a pool compare?

45. FILE Refer to the Baseball 2016 data that report information on the 30 Major League Baseball teams for the 2016 season.
- **a.** In the data set, the year opened, is the first year of operation for that stadium. For each team, use this variable to create a new variable, stadium age, by subtracting the value of the variable, year opened, from the current year. Develop a box plot with the new variable, age. Are there any outliers? If so, which of the stadiums are outliers?
- **b.** Using the variable, salary, create a box plot. Are there any outliers? Compute the quartiles using formula (4–1). Write a brief summary of your analysis.
- **c.** Draw a scatter diagram with the variable, wins, on the vertical axis and salary on the horizontal axis. What are your conclusions?
- **d.** Using the variable, wins, draw a dot plot. What can you conclude from this plot?

46. FILE Refer to the Lincolnville School District bus data.
- **a.** Referring to the maintenance cost variable, develop a box plot. What are the minimum, first quartile, median, third quartile, and maximum values? Are there any outliers?
- **b.** Using the median maintenance cost, develop a contingency table with bus manufacturer as one variable and whether the maintenance cost was above or below the median as the other variable. What are your conclusions?

A REVIEW OF CHAPTERS 1–4

This section is a review of the major concepts and terms introduced in Chapters 1–4. Chapter 1 began by describing the meaning and purpose of statistics. Next we described the different types of variables and the four levels of measurement. Chapter 2 was concerned with describing a set of observations by organizing it into a frequency distribution and then portraying the frequency distribution as a histogram or a frequency polygon. Chapter 3 began by describing measures of location, such as the mean, weighted mean, median, geometric mean, and mode. This chapter also included measures of dispersion, or spread. Discussed in this section were the range, variance, and standard deviation. Chapter 4 included several graphing techniques such as dot plots, box plots, and scatter diagrams. We also discussed the coefficient of skewness, which reports the lack of symmetry in a set of data.

Throughout this section we stressed the importance of statistical software, such as Excel and Minitab. Many computer outputs in these chapters demonstrated how quickly and effectively a large data set can be organized into a frequency distribution, several of the measures of location or measures of variation calculated, and the information presented in graphical form.

PROBLEMS

1. FILE The duration in minutes of a sample of 50 power outages last year in the state of South Carolina is listed below.

124	14	150	289	52	156	203	82	27	248
39	52	103	58	136	249	110	298	251	157
186	107	142	185	75	202	119	219	156	78
116	152	206	117	52	299	58	153	219	148
145	187	165	147	158	146	185	186	149	140

Use a statistical software package such as Excel or Minitab to help answer the following questions.

a. Determine the mean, median, and standard deviation.
b. Determine the first and third quartiles.
c. Develop a box plot. Are there any outliers? Do the amounts follow a symmetric distribution or are they skewed? Justify your answer.
d. Organize the distribution of funds into a frequency distribution.
e. Write a brief summary of the results in parts a to d.

2. FILE Listed below are the 45 U.S. presidents and their age as they began their terms in office.

Number	Name	Age	Number	Name	Age
1	Washington	57	24	Cleveland	55
2	J. Adams	61	25	McKinley	54
3	Jefferson	57	26	T. Roosevelt	42
4	Madison	57	27	Taft	51
5	Monroe	58	28	Wilson	56
6	J. Q. Adams	57	29	Harding	55
7	Jackson	61	30	Coolidge	51
8	Van Buren	54	31	Hoover	54
9	W. H. Harrison	68	32	F. D. Roosevelt	51
10	Tyler	51	33	Truman	60
11	Polk	49	34	Eisenhower	62
12	Taylor	64	35	Kennedy	43
13	Fillmore	50	36	L. B. Johnson	55
14	Pierce	48	37	Nixon	56
15	Buchanan	65	38	Ford	61
16	Lincoln	52	39	Carter	52
17	A. Johnson	56	40	Reagan	69
18	Grant	46	41	G. H. W. Bush	64
19	Hayes	54	42	Clinton	46
20	Garfield	49	43	G. W. Bush	54
21	Arthur	50	44	Obama	47
22	Cleveland	47	45	Trump	70
23	B. Harrison	55			

Use a statistical software package such as Excel or Minitab to help answer the following questions.

a. Determine the mean, median, and standard deviation.
b. Determine the first and third quartiles.
c. Develop a box plot. Are there any outliers? Do the amounts follow a symmetric distribution or are they skewed? Justify your answer.
d. Organize the distribution of ages into a frequency distribution.
e. Write a brief summary of the results in parts a to d.

3. FILE Listed below is the 2014 median household income for the 50 states and the District of Columbia. https://www.census.gov/hhes/www/income/data/historical/household/

State	Amount	State	Amount
Alabama	42,278	Montana	51,102
Alaska	67,629	Nebraska	56,870
Arizona	49,254	Nevada	49,875
Arkansas	44,922	New Hampshire	73,397
California	60,487	New Jersey	65,243
Colorado	60,940	New Mexico	46,686
Connecticut	70,161	New York	54,310
Delaware	57,522	North Carolina	46,784
D.C.	68,277	North Dakota	60,730
Florida	46,140	Ohio	49,644
Georgia	49,555	Oklahoma	47,199
Hawaii	71,223	Oregon	58,875
Idaho	53,438	Pennsylvania	55,173
Illinois	54,916	Rhode Island	58,633
Indiana	48,060	South Carolina	44,929
Iowa	57,810	South Dakota	53,053
Kansas	53,444	Tennessee	43,716
Kentucky	42,786	Texas	53,875
Louisiana	42,406	Utah	63,383
Maine	51,710	Vermont	60,708
Maryland	76,165	Virginia	66,155
Massachusetts	63,151	Washington	59,068
Michigan	52,005	West Virginia	39,552
Minnesota	67,244	Wisconsin	58,080
Mississippi	35,521	Wyoming	55,690
Missouri	56,630		

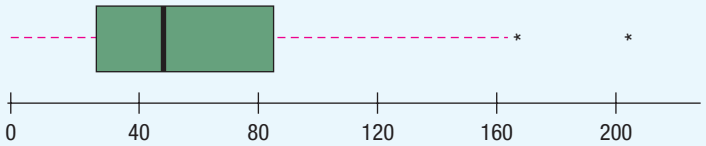
Use a statistical software package such as Excel or Minitab to help answer the following questions.

a. Determine the mean, median, and standard deviation.
b. Determine the first and third quartiles.
c. Develop a box plot. Are there any outliers? Do the amounts follow a symmetric distribution or are they skewed? Justify your answer.
d. Organize the distribution of funds into a frequency distribution.
e. Write a brief summary of the results in parts a to d.

4. A sample of 12 homes sold last week in St. Paul, Minnesota, revealed the following information. Draw a scatter diagram. Can we conclude that, as the size of the home (reported below in thousands of square feet) increases, the selling price (reported in $ thousands) also increases?

Home Size (thousands of square feet)	Selling Price ($ thousands)	Home Size (thousands of square feet)	Selling Price ($ thousands)
1.4	100	1.3	110
1.3	110	0.8	85
1.2	105	1.2	105
1.1	120	0.9	75
1.4	80	1.1	70
1.0	105	1.1	95

5. Refer to the following diagram.



a. What is the graph called?
b. What are the median, and first and third quartile values?
c. Is the distribution positively skewed? Tell how you know.
d. Are there any outliers? If yes, estimate these values.
e. Can you determine the number of observations in the study?

CASES

A. Century National Bank

The following case will appear in subsequent review sections. Assume that you work in the Planning Department of the Century National Bank and report to Ms. Lamberg. You will need to do some data analysis and prepare a short written report. Remember, Mr. Selig is the president of the bank, so you will want to ensure that your report is complete and accurate. A copy of the data appears in Appendix A.6.

Century National Bank has offices in several cities in the Midwest and the southeastern part of the United States. Mr. Dan Selig, president and CEO, would like to know the characteristics of his checking account customers. What is the balance of a typical customer?

How many other bank services do the checking account customers use? Do the customers use the ATM service and, if so, how often? What about debit cards? Who uses them, and how often are they used?

To better understand the customers, Mr. Selig asked Ms. Wendy Lamberg, director of planning, to select a sample of customers and prepare a report. To begin, she has appointed a team from her staff. You are the head of the team and responsible for preparing the report. You select a random sample of 60 customers. In addition to the balance in each account at the end of last month, you determine (1) the number of ATM (automatic teller machine) transactions in the last month; (2) the number of other bank services (a savings account, a certificate of deposit, etc.) the customer uses; (3) whether the customer has a debit card (this is a bank service in which charges are made directly to the customer's account); and (4) whether or not interest is paid on the checking account. The sample includes customers from the branches in Cincinnati, Ohio; Atlanta, Georgia; Louisville, Kentucky; and Erie, Pennsylvania.

1. Develop a graph or table that portrays the checking balances. What is the balance of a typical customer? Do many customers have more than $2,000 in their accounts? Does it appear that there is a difference in the distribution of the accounts among the four branches? Around what value do the account balances tend to cluster?

2. Determine the mean and median of the checking account balances. Compare the mean and the median balances for the four branches. Is there a difference among the branches? Be sure to explain the difference between the mean and the median in your report.

3. Determine the range and the standard deviation of the checking account balances. What do the first and third quartiles show? Determine the coefficient of skewness and indicate what it shows. Because Mr. Selig does not deal with statistics daily, include a brief description and interpretation of the standard deviation and other measures.

B. Wildcat Plumbing Supply Inc.: Do We Have Gender Differences?

Wildcat Plumbing Supply has served the plumbing needs of Southwest Arizona for more than 40 years. The company was founded by Mr. Terrence St. Julian and is run today by his son Cory. The company has grown from a handful of employees to more than 500 today. Cory is concerned about several positions within the company where he has men and women doing essentially the same job but at different pay. To investigate, he collected the information below. Suppose you are a student intern in the Accounting Department and have been given the task to write a report summarizing the situation.

Yearly Salary ($000)	Women	Men
Less than 30	2	0
30 up to 40	3	1
40 up to 50	17	4
50 up to 60	17	24
60 up to 70	8	21
70 up to 80	3	7
80 or more	0	3

To kick off the project, Mr. Cory St. Julian held a meeting with his staff and you were invited. At this meeting, it was suggested that you calculate several measures of

location, create charts or draw graphs such as a cumulative frequency distribution, and determine the quartiles for both men and women. Develop the charts and write the report summarizing the yearly salaries of employees at Wildcat Plumbing Supply. Does it appear that there are pay differences based on gender?

C. Kimble Products: Is There a Difference In the Commissions?

At the January national sales meeting, the CEO of Kimble Products was questioned extensively regarding the company policy for paying commissions to its sales representatives. The company sells sporting goods to two major markets. There are 40 sales representatives who call directly on large-volume customers, such as the athletic departments at major colleges and universities and professional sports franchises. There are 30 sales representatives who represent the company to retail stores located in shopping malls and large discounters such as Kmart and Target.

Upon his return to corporate headquarters, the CEO asked the sales manager for a report comparing the commissions earned last year by the two parts of the sales team. The information is reported below. Write a brief report. Would you conclude that there is a difference? Be sure to include information in the report on both the central tendency and dispersion of the two groups.

Commissions Earned by Sales Representatives Calling on Athletic Departments ($)

354	87	1,676	1,187	69	3,202	680	39	1,683	1,106
883	3,140	299	2,197	175	159	1,105	434	615	149
1,168	278	579	7	357	252	1,602	2,321	4	392
416	427	1,738	526	13	1,604	249	557	635	527

Commissions Earned by Sales Representatives Calling on Large Retailers ($)

1,116	681	1,294	12	754	1,206	1,448	870	944	1,255
1,213	1,291	719	934	1,313	1,083	899	850	886	1,556
886	1,315	1,858	1,262	1,338	1,066	807	1,244	758	918

PRACTICE TEST

There is a practice test at the end of each review section. The tests are in two parts. The first part contains several objective questions, usually in a fill-in-the-blank format. The second part is problems. In most cases, it should take 30 to 45 minutes to complete the test. The problems require a calculator. Check the answers in the Answer Section in the back of the book.

Part 1—Objective

1. The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making effective decisions is called _____. 1. _____
2. Methods of organizing, summarizing, and presenting data in an informative way are called _____. 2. _____
3. The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest are called the _____. 3. _____
4. List the two types of variables. 4. _____
5. The number of bedrooms in a house is an example of a _____. (discrete variable, continuous variable, qualitative variable—pick one) 5. _____
6. The jersey numbers of Major League Baseball players are an example of what level of measurement? 6. _____
7. The classification of students by eye color is an example of what level of measurement? 7. _____
8. The sum of the differences between each value and the mean is always equal to what value? 8. _____
9. A set of data contained 70 observations. How many classes would the 2^k method suggest to construct a frequency distribution? 9. _____
10. What percent of the values in a data set are always larger than the median? 10. _____
11. The square of the standard deviation is the _____. 11. _____
12. The standard deviation assumes a negative value when _____. (all the values are negative, at least half the values are negative, or never—pick one.) 12. _____
13. Which of the following is least affected by an outlier? (mean, median, or range—pick one) 13. _____

Part 2—Problems

1. The Russell 2000 index of stock prices increased by the following amounts over the last 3 years.

18%	4%	2%

What is the geometric mean increase for the 3 years? _____

2. The information below refers to the selling prices ($000) of homes sold in Warren, Pennsylvania, during 2016.

Selling Price ($000)	Frequency
120.0 up to 150.0	4
150.0 up to 180.0	18
180.0 up to 210.0	30
210.0 up to 240.0	20
240.0 up to 270.0	17
270.0 up to 300.0	10
300.0 up to 330.0	6

a. What is the class interval? _____
b. How many homes were sold in 2016? _____
c. How many homes sold for less than $210,000? _____
d. What is the relative frequency of the 210 up to 240 class? _____
e. What is the midpoint of the 150 up to 180 class? _____
f. The selling prices range between what two amounts? _____

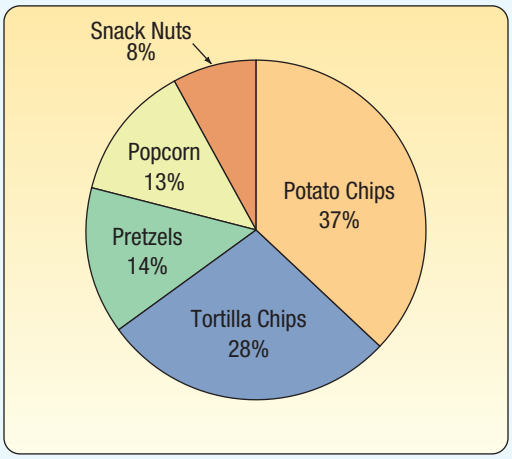
3. A sample of eight college students revealed they owned the following number of CDs.

52 76 64 79 80 74 66 69

a. What is the mean number of CDs owned? _____
b. What is the median number of CDs owned? _____
c. What is the 40th percentile? _____
d. What is the range of the number of CDs owned? _____
e. What is the standard deviation of the number of CDs owned? _____

4. An investor purchased 200 shares of the Blair Company for $36 each in July of 2013, 300 shares at $40 each in September 2015, and 500 shares at $50 each in January 2016. What is the investor's weighted mean price per share? _____

5. During the 50th Super Bowl, 30 million pounds of snack food were eaten. The chart below depicts this information.



a. What is the name given to this graph? _____
b. Estimate, in millions of pounds, the amount of potato chips eaten during the game. _____
c. Estimate the relationship of potato chips to popcorn. (twice as much, half as much, three times, none of these—pick one) _____
d. What percent of the total do potato chips and tortilla chips comprise? _____

5 A Survey of Probability Concepts

© Karin Slade/Getty Image

▲ **RECENT SURVEYS** indicate 60% of tourists to China visited the Forbidden City, the Temple of Heaven, the Great Wall, and other historical sites in or near Beijing. Forty percent visited Xi'an and its magnificent terra-cotta soldiers, horses, and chariots, which lay buried for over 2,000 years. Thirty percent of the tourists went to both Beijing and Xi'an. What is the probability that a tourist visited at least one of these places? (See Exercise 76 and LO5-3.)

LEARNING OBJECTIVES

When you have completed this chapter, you will be able to:

LO5-1 Define the terms *probability, experiment, event,* and *outcome.*

LO5-2 Assign probabilities using a classical, empirical, or subjective approach.

LO5-3 Calculate probabilities using the rules of addition.

LO5-4 Calculate probabilities using the rules of multiplication.

LO5-5 Compute probabilities using a contingency table.

LO5-6 Calculate probabilities using Bayes' theorem.

LO5-7 Determine the number of outcomes using principles of counting.

INTRODUCTION

The emphasis in Chapters 2, 3, and 4 is on descriptive statistics. In Chapter 2, we organize the profits on 180 vehicles sold by the Applewood Auto Group into a frequency distribution. This frequency distribution shows the smallest and the largest profits and where the largest concentration of data occurs. In Chapter 3, we use numerical measures of location and dispersion to locate a typical profit on vehicle sales and to examine the variation in the profit of a sale. We describe the variation in the profits with such measures of dispersion as the range and the standard deviation. In Chapter 4, we develop charts and graphs, such as a scatter diagram or a dot plot, to further describe the data graphically.

Descriptive statistics is concerned with summarizing data collected from past events. We now turn to the second facet of statistics, namely, *computing the chance that something will occur in the future*. This facet of statistics is called **statistical inference** or **inferential statistics.**

Seldom does a decision maker have complete information to make a decision. For example:

© Ballda/Shutterstock.com

- Toys and Things, a toy and puzzle manufacturer, recently developed a new game based on sports trivia. It wants to know whether sports buffs will purchase the game. "Slam Dunk" and "Home Run" are two of the names under consideration. To investigate, the president of Toys and Things decided to hire a market research firm. The firm selected a sample of 800 consumers from the population and asked each respondent for a reaction to the new game and its proposed titles. Using the sample results, the company can estimate the proportion of the population that will purchase the game.
- The quality assurance department of a U.S. Steel mill must assure management that the quarter-inch wire being produced has an acceptable tensile strength. Clearly, not all the wire produced can be tested for tensile strength because testing requires the wire to be stretched until it breaks—thus destroying it. So a random sample of 10 pieces is selected and tested. Based on the test results, all the wire produced is deemed to be either acceptable or unacceptable.
- Other questions involving uncertainty are: Should the daytime drama *Days of Our Lives* be discontinued immediately? Will a newly developed mint-flavored cereal be profitable if marketed? Will Charles Linden be elected to county auditor in Batavia County?

Statistical inference deals with conclusions about a population based on a sample taken from that population. (The populations for the preceding illustrations are all consumers who like sports trivia games, all the quarter-inch steel wire produced, all television viewers who watch soaps, all who purchase breakfast cereal, and so on.)

Because there is uncertainty in decision making, it is important that all the known risks involved be scientifically evaluated. Helpful in this evaluation is *probability theory,* often referred to as the science of uncertainty. Probability theory allows the decision maker to analyze the risks and minimize the gamble inherent, for example, in marketing a new product or accepting an incoming shipment possibly containing defective parts.

Because probability concepts are so important in the field of statistical inference (to be discussed starting with Chapter 8), this chapter introduces the basic language of probability, including such terms as *experiment, event, subjective probability,* and *addition* and *multiplication rules*.

STATISTICS IN ACTION

Government statistics show there are about 1.7 automobile-caused fatalities for every 100,000,000 vehicle-miles. If you drive 1 mile to the store to buy your lottery ticket and then return home, you have driven 2 miles. Thus the probability that you will join this statistical group on your next 2-mile round trip is $2 \times 1.7/100{,}000{,}000 = 0.000000034$. This can also be stated as "One in 29,411,765." Thus, if you drive to the store to buy your Powerball ticket, your chance of being killed (or killing someone else) is more than 4 times greater than the chance that you will win the Powerball Jackpot, one chance in 120,526,770.
http://www.durangobill.com/PowerballOdds.html

LO5-1
Define the terms *probability, experiment, event,* and *outcome*.

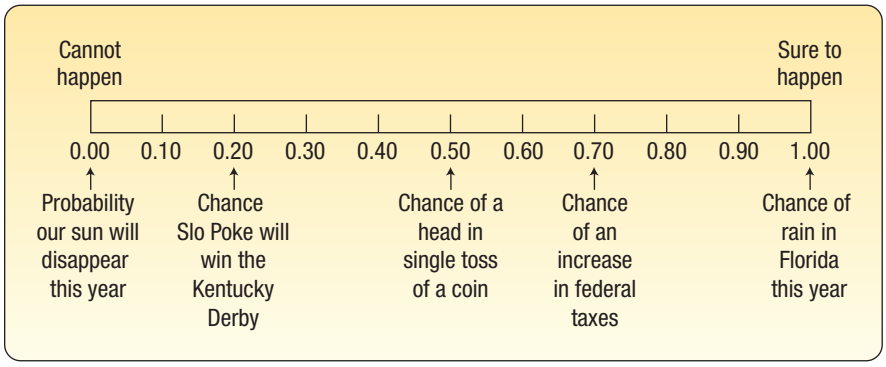
WHAT IS A PROBABILITY?

No doubt you are familiar with terms such as *probability, chance,* and *likelihood*. They are often used interchangeably. The weather forecaster announces that there is a 70% chance of rain for Super Bowl Sunday. Based on a survey of consumers who tested a newly developed toothpaste with a banana flavor, the probability is .03 that, if marketed, it will be a financial success. (This means that the chance of the banana-flavor toothpaste being accepted by the public is rather remote.) What is a **probability?** In general, it is a numerical value that describes the chance that something will happen.

PROBABILITY A value between zero and one, inclusive, describing the relative possibility (chance or likelihood) an event will occur.

A probability is frequently expressed as a decimal, such as .70, .27, or .50, or a percent such as 70%, 27% or 50%. It also may be reported as a fraction such as 7/10, 27/100, or 1/2. It can assume any number from 0 to 1, inclusive. Expressed as a percentage, the range is between 0% and 100%, inclusive. If a company has only five sales regions, and each region's name or number is written on a slip of paper and the slips put in a hat, the probability of selecting one of the five regions is 1. The probability of selecting from the hat a slip of paper that reads "Pittsburgh Steelers" is 0. Thus, the probability of 1 represents something that is certain to happen, and the probability of 0 represents something that cannot happen.

The closer a probability is to 0, the more improbable it is the event will happen. The closer the probability is to 1, the more likely it will happen. The relationship is shown in the following diagram along with a few of our personal beliefs. You might, however, select a different probability for Slo Poke's chances to win the Kentucky Derby or for an increase in federal taxes.



Sometimes, the likelihood of an event is expressed using the term *odds*. To explain, someone says the odds are "five to two" that an event will occur. This means that in a total of seven trials (5 + 2), the event will occur five times and not occur two times. Using odds, we can compute the probability that the event occurs as 5/(5 + 2) or 5/7. So, if the odds in favor of an event are x to y, the probability of the event is $x/(x + y)$.

Three key words are used in the study of probability: **experiment, outcome,** and **event.** These terms are used in our everyday language, but in statistics they have specific meanings.

EXPERIMENT A process that leads to the occurrence of one and only one of several possible results.

This definition is more general than the one used in the physical sciences, where we picture someone manipulating test tubes or microscopes. In reference to probability, an experiment has two or more possible results, and it is uncertain which will occur.

OUTCOME A particular result of an experiment.

For example, the tossing of a coin is an experiment. You are unsure of the outcome. When a coin is tossed, one particular outcome is a "head." The alternative outcome is a "tail." Similarly, asking 500 college students if they would travel more than 100 miles to attend a Mumford and Sons concert is an experiment. In this experiment, one possible outcome is that 273 students indicate they would travel more than 100 miles to attend the concert. Another outcome is that 317 students would attend the concert. Still another outcome is that 423 students indicate they would attend the concert. When one or more of the experiment's outcomes are observed, we call this an event.

EVENT A collection of one or more outcomes of an experiment.

Examples to clarify the definitions of the terms *experiment, outcome,* and *event* are presented in the following figure.

In the die-rolling experiment, there are six possible outcomes, but there are many possible events. When counting the number of members of the board of directors for Fortune 500 companies over 60 years of age, the number of possible outcomes can be anywhere from zero to the total number of members. There are an even larger number of possible events in this experiment.

Experiment	Roll a die	Count the number of members of the board of directors for Fortune 500 companies who are over 60 years of age
All possible outcomes	Observe a 1 Observe a 2 Observe a 3 Observe a 4 Observe a 5 Observe a 6	None is over 60 One is over 60 Two are over 60 ... 29 are over 60 ... ... 48 are over 60 ...
Some possible events	Observe an even number Observe a number greater than 4 Observe a number 3 or less	More than 13 are over 60 Fewer than 20 are over 60

SELF-REVIEW 5–1

Video Games Inc. recently developed a new video game. Its playability is to be tested by 80 veteran game players.

(a) What is the experiment?
(b) What is one possible outcome?
(c) Suppose 65 of the 80 players testing the new game said they liked it. Is 65 a probability?
(d) The probability that the new game will be a success is computed to be −1.0. Comment.
(e) Specify one possible event.

LO5-2
Assign probabilities using a classical, empirical, or subjective approach.

APPROACHES TO ASSIGNING PROBABILITIES

In this section, we describe three ways to assign a probability to an event: classical, empirical, and subjective. The classical and empirical methods are objective and are based on information and data. The subjective method is based on a person's belief or estimate of an event's likelihood.

Classical Probability

Classical probability is based on the assumption that the outcomes of an experiment are *equally likely*. Using the classical viewpoint, the probability of an event happening is computed by dividing the number of favorable outcomes by the number of possible outcomes:

CLASSICAL PROBABILITY

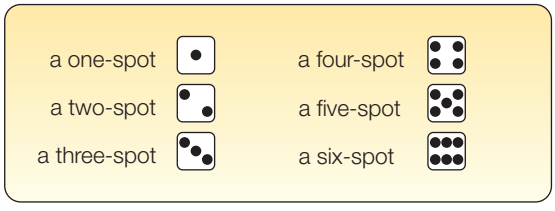
$$\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} \quad [5\text{–}1]$$

EXAMPLE

Consider an experiment of rolling a six-sided die. What is the probability of the event "an even number of spots appear face up"?

SOLUTION

The possible outcomes are:



There are three "favorable" outcomes (a two, a four, and a six) in the collection of six equally likely possible outcomes. Therefore:

$$\text{Probability of an even number} = \frac{3}{6} = .5 \quad \leftarrow \quad \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

The **mutually exclusive** concept appeared earlier in our study of frequency distributions in Chapter 2. Recall that we create classes so that a particular value is included in only one of the classes and there is no overlap between classes. Thus, only one of several events can occur at a particular time.

MUTUALLY EXCLUSIVE The occurrence of one event means that none of the other events can occur at the same time.

The variable "gender" presents mutually exclusive outcomes, male and female. An employee selected at random is either male or female but cannot be both. A manufactured part is acceptable or unacceptable. The part cannot be both acceptable and unacceptable at the same time. In a sample of manufactured parts, the event of selecting an unacceptable part and the event of selecting an acceptable part are mutually exclusive.

If an experiment has a set of events that includes every possible outcome, such as the events "an even number" and "an odd number" in the die-tossing experiment, then the set of events is **collectively exhaustive.** For the die-tossing experiment, every outcome will be either even or odd. So the set is collectively exhaustive.

COLLECTIVELY EXHAUSTIVE At least one of the events must occur when an experiment is conducted.

If the set of events is collectively exhaustive and the events are mutually exclusive, the sum of the probabilities is 1. Historically, the classical approach to probability was developed and applied in the 17th and 18th centuries to games of chance, such as cards and dice. It is unnecessary to do an experiment to determine the probability of an event occurring using the classical approach because the total number of outcomes is known before the experiment. The flip of a coin has two possible outcomes; the roll of a die has six possible outcomes. We can logically arrive at the probability of getting a tail on the toss of one coin or three heads on the toss of three coins.

The classical approach to probability can also be applied to lotteries. In South Carolina, one of the games of the Education Lottery is "Pick 3." A person buys a lottery ticket and selects three numbers between 0 and 9. Once per week, the three numbers are randomly selected from a machine that tumbles three containers each with balls numbered 0 through 9. One way to win is to match the numbers and the order of the numbers. Given that 1,000 possible outcomes exist (000 through 999), the probability of winning with any three-digit number is 0.001, or 1 in 1,000.

Empirical Probability

Empirical or relative frequency is the second type of objective probability. It is based on the number of times an event occurs as a proportion of a known number of trials.

EMPIRICAL PROBABILITY The probability of an event happening is the fraction of the time similar events happened in the past.

The formula to determine an empirical probability is:

$$\text{Empirical probability} = \frac{\text{Number of times the event occurs}}{\text{Total number of observations}}$$

The empirical approach to probability is based on what is called the **law of large numbers.** The key to establishing probabilities empirically is that more observations will provide a more accurate estimate of the probability.

LAW OF LARGE NUMBERS Over a large number of trials, the empirical probability of an event will approach its true probability.

To explain the law of large numbers, suppose we toss a fair coin. The result of each toss is either a head or a tail. With just one toss of the coin the empirical probability for

heads is either zero or one. If we toss the coin a great number of times, the probability of the outcome of heads will approach .5. The following table reports the results of seven different experiments of flipping a fair coin 1, 10, 50, 100, 500, 1,000, and 10,000 times and then computing the relative frequency of heads. Note as we increase the number of trials, the empirical probability of a head appearing approaches .5, which is its value based on the classical approach to probability.

Number of Trials	Number of Heads	Relative Frequency of Heads
1	0	.00
10	3	.30
50	26	.52
100	52	.52
500	236	.472
1,000	494	.494
10,000	5,027	.5027

What have we demonstrated? Based on the classical definition of probability, the likelihood of obtaining a head in a single toss of a fair coin is .5. Based on the empirical or relative frequency approach to probability, the probability of the event happening approaches the same value based on the classical definition of probability.

This reasoning allows us to use the empirical or relative frequency approach to finding a probability. Here are some examples.

- Last semester, 80 students registered for Business Statistics 101 at Scandia University. Twelve students earned an A. Based on this information and the empirical approach to assigning a probability, we estimate the likelihood a student at Scandia will earn an A is .15.
- Stephen Curry of the Golden State Warriors made 363 out of 400 free throw attempts during the 2015–16 NBA season. Based on the empirical approach to probability, the likelihood of him making his next free throw attempt is .908.

Life insurance companies rely on past data to determine the acceptability of an applicant as well as the premium to be charged. Mortality tables list the likelihood a person of a particular age will die within the upcoming year. For example, the likelihood a 20-year-old female will die within the next year is .00105.

The empirical concept is illustrated with the following example.

EXAMPLE

On February 1, 2003, the Space Shuttle *Columbia* exploded. This was the second disaster in 113 space missions for NASA. On the basis of this information, what is the probability that a future mission is successfully completed?

SOLUTION

We use letters or numbers to simplify the equations. P stands for probability and A represents the event of a successful mission. In this case, $P(A)$ stands for the probability a future mission is successfully completed.

$$\text{Probability of a successful flight} = \frac{\text{Number of successful flights}}{\text{Total number of flights}}$$

$$P(A) = \frac{111}{113} = .98$$

We can use this as an estimate of probability. In other words, based on past experience, the probability is .98 that a future space shuttle mission will be safely completed.

Subjective Probability

If there is little or no experience or information on which to base a probability, it is estimated subjectively. Essentially, this means an individual evaluates the available opinions and information and then estimates or assigns the probability. This probability is called a **subjective probability.**

SUBJECTIVE CONCEPT OF PROBABILITY The likelihood (probability) of a particular event happening that is assigned by an individual based on whatever information is available.

Illustrations of subjective probability are:

1. Estimating the likelihood the New England Patriots will play in the Super Bowl next year.
2. Estimating the likelihood you are involved in an automobile accident during the next 12 months.
3. Estimating the likelihood the U.S. budget deficit will be reduced by half in the next 10 years.

The types of probability are summarized in Chart 5–1. A probability statement always assigns a likelihood to an event that has not yet occurred. There is, of course, considerable latitude in the degree of uncertainty that surrounds this probability, based primarily on the knowledge possessed by the individual concerning the underlying process. The individual possesses a great deal of knowledge about the toss of a die and can state that the probability that a one-spot will appear face up on the toss of a true die is one-sixth. But we know very little concerning the acceptance in the marketplace of a new and untested product. For example, even though a market research director tests a newly developed product in 40 retail stores and states that there is a 70% chance that the product will have sales of more than 1 million units, she has limited knowledge of how consumers will react when it is marketed nationally. In both cases (the case of the person rolling a die and the testing of a new product), the individual is assigning a probability value to an event of interest, and a difference exists only in the predictor's confidence in the precision of the estimate. However, regardless of the viewpoint, the same laws of probability (presented in the following sections) will be applied.

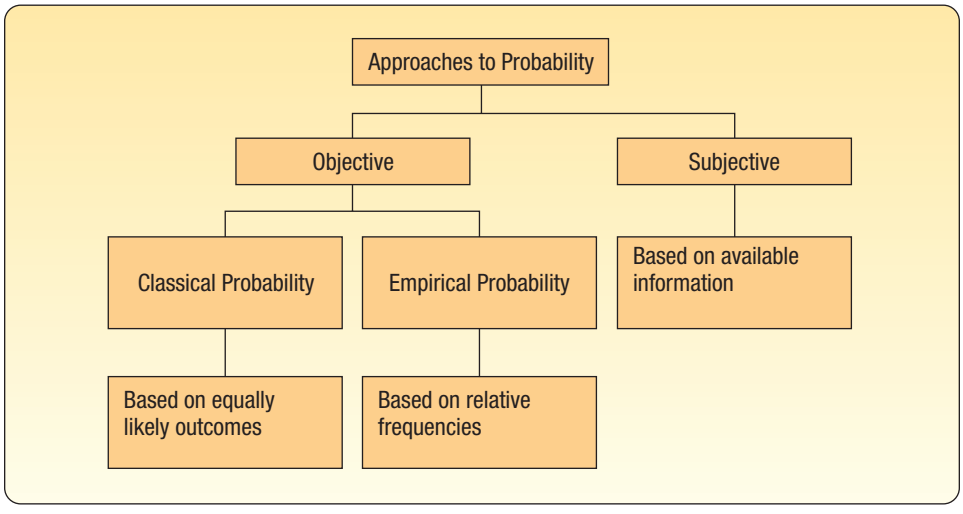


CHART 5–1 Summary of Approaches to Probability

SELF-REVIEW 5–2

1. One card will be randomly selected from a standard 52-card deck. What is the probability the card will be a queen? Which approach to probability did you use to answer this question?
2. The Center for Child Care reports on 539 children and the marital status of their parents. There are 333 married, 182 divorced, and 24 widowed parents. What is the probability a particular child chosen at random will have a parent who is divorced? Which approach did you use?
3. What is the probability you will save one million dollars by the time you retire? Which approach to probability did you use to answer this question?

EXERCISES

1. Some people are in favor of reducing federal taxes to increase consumer spending and others are against it. Two persons are selected and their opinions are recorded. Assuming no one is undecided, list the possible outcomes.

2. A quality control inspector selects a part to be tested. The part is then declared acceptable, repairable, or scrapped. Then another part is tested. List the possible outcomes of this experiment regarding two parts.

3. FILE A survey of 34 students at the Wall College of Business showed the following majors:

Accounting	10
Finance	5
Economics	3
Management	6
Marketing	10

From the 34 students, suppose you randomly select a student.
- **a.** What is the probability he or she is a management major?
- **b.** Which concept of probability did you use to make this estimate?

4. A large company must hire a new president. The Board of Directors prepares a list of five candidates, all of whom are equally qualified. Two of these candidates are members of a minority group. To avoid bias in the selection of the candidate, the company decides to select the president by lottery.
- **a.** What is the probability one of the minority candidates is hired?
- **b.** Which concept of probability did you use to make this estimate?

5. In each of the following cases, indicate whether classical, empirical, or subjective probability is used.
- **a.** A baseball player gets a hit in 30 out of 100 times at bat. The probability is .3 that he gets a hit in his next at bat.
- **b.** A seven-member committee of students is formed to study environmental issues. What is the likelihood that any one of the seven is randomly chosen as the spokesperson?
- **c.** You purchase a ticket for the Lotto Canada lottery. Over 5 million tickets were sold. What is the likelihood you will win the $1 million jackpot?
- **d.** The probability of an earthquake in northern California in the next 10 years above 5.0 on the Richter Scale is .80.

6. A firm will promote two employees out of a group of six men and three women.
- **a.** List all possible outcomes.
- **b.** What probability concept would be used to assign probabilities to the outcomes?

7. A sample of 40 oil industry executives was selected to test a questionnaire. One question about environmental issues required a yes or no answer.
- **a.** What is the experiment?
- **b.** List one possible event.
- **c.** Ten of the 40 executives responded yes. Based on these sample responses, what is the probability that an oil industry executive will respond yes?
- **d.** What concept of probability does this illustrate?
- **e.** Are each of the possible outcomes equally likely and mutually exclusive?

8. FILE A sample of 2,000 licensed drivers revealed the following number of speeding violations.

Number of Violations	Number of Drivers
0	1,910
1	46
2	18
3	12
4	9
5 or more	5
Total	2,000

a. What is the experiment?
b. List one possible event.
c. What is the probability that a particular driver had exactly two speeding violations?
d. What concept of probability does this illustrate?

9. Bank of America customers select their own three-digit personal identification number (PIN) for use at ATMs.
a. Think of this as an experiment and list four possible outcomes.
b. What is the probability that a customer will pick 259 as their PIN?
c. Which concept of probability did you use to answer (b)?

10. An investor buys 100 shares of AT&T stock and records its price change daily.
a. List several possible events for this experiment.
b. Which concept of probability did you use in (a)?

LO5-3
Calculate probabilities using the rules of addition.

RULES OF ADDITION FOR COMPUTING PROBABILITIES

There are two rules of addition, the special rule of addition and the general rule of addition. We begin with the special rule of addition.

Special Rule of Addition

When we use the **special rule of addition,** the events must be *mutually exclusive*. Recall that mutually exclusive means that when one event occurs, none of the other events can occur at the same time. An illustration of mutually exclusive events in the die-tossing experiment is the events "a number 4 or larger" and "a number 2 or smaller." If the outcome is in the first group {4, 5, and 6}, then it cannot also be in the second group {1 and 2}. Another illustration is a product coming off the assembly line cannot be defective and satisfactory at the same time.

If two events A and B are mutually exclusive, the special rule of addition states that the probability of one *or* the other event's occurring equals the sum of their probabilities. This rule is expressed in the following formula:

SPECIAL RULE OF ADDITION $P(A \text{ or } B) = P(A) + P(B)$ **[5–2]**

For three mutually exclusive events designated A, B, and C, the rule is written:

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

An example will show the details.

EXAMPLE

© Ian Dagnall/Alamy Stock Photo

A machine fills plastic bags with a mixture of beans, broccoli, and other vegetables. Most of the bags contain the correct weight, but because of the variation in the size of the beans and other vegetables, a package might be underweight or overweight. A check of 4,000 packages filled in the past month revealed:

Weight	Event	Number of Packages	Probability of Occurrence	
Underweight	*A*	100	.025	← 100 / 4,000
Satisfactory	*B*	3,600	.900	
Overweight	*C*	300	.075	
		4,000	1.000	

What is the probability that a particular package will be either underweight or overweight?

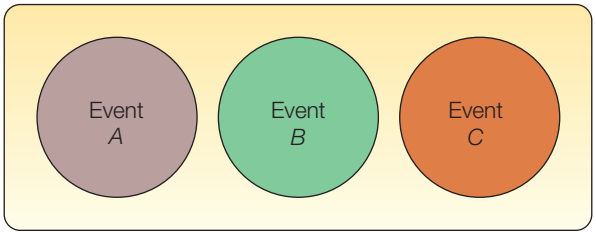
SOLUTION

The outcome "underweight" is the event *A*. The outcome "overweight" is the event *C*. Applying the special rule of addition:

$$P(A \text{ or } C) = P(A) + P(C) = .025 + .075 = .10$$

Note that the events are mutually exclusive, meaning that a package of mixed vegetables cannot be underweight, satisfactory, and overweight at the same time. They are also collectively exhaustive; that is, a selected package must be either underweight, satisfactory, or overweight.

English logician J. Venn (1834–1923) developed a diagram to portray graphically the outcome of an experiment. The *mutually exclusive* concept and various other rules for combining probabilities can be illustrated using this device. To construct a Venn diagram, a space is first enclosed representing the total of all possible outcomes. This space is usually in the form of a rectangle. An event is then represented by a circular area that is drawn inside the rectangle proportional to the probability of the event. The following Venn diagram represents the *mutually exclusive* concept. There is no overlapping of events, meaning that the events are mutually exclusive. In the following Venn diagram, assume the events *A*, *B*, and *C* are about equally likely.



Complement Rule

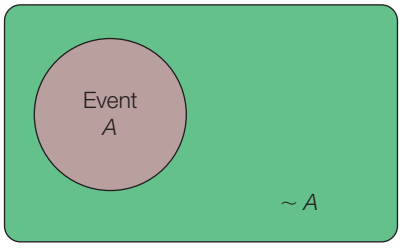
The probability that a bag of mixed vegetables selected is underweight, $P(A)$, plus the probability that it is not an underweight bag, written $P(\sim A)$ and read "not A," must logically equal 1. This is written:

$$P(A) + P(\sim A) = 1$$

This can be revised to read:

COMPLEMENT RULE $P(A) = 1 - P(\sim A)$ **[5–3]**

This is the **complement rule.** It is used to determine the probability of an event occurring by subtracting the probability of the event not occurring from 1. This rule is useful because sometimes it is easier to calculate the probability of an event happening by determining the probability of it not happening and subtracting the result from 1. Notice that the events A and $\sim A$ are mutually exclusive and collectively exhaustive. Therefore, the probabilities of A and $\sim A$ sum to 1. A Venn diagram illustrating the complement rule is shown as:

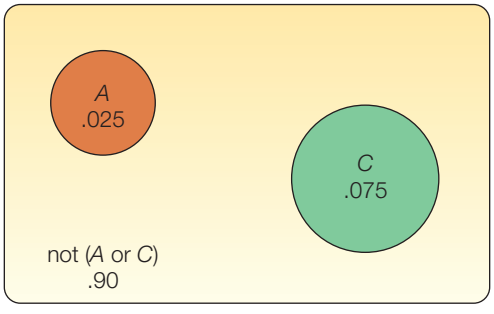


EXAMPLE

Referring to the previous example/solution, the probability a bag of mixed vegetables is underweight is .025 and the probability of an overweight bag is .075. Use the complement rule to show the probability of a satisfactory bag is .900. Show the solution using a Venn diagram.

SOLUTION

The probability the bag is unsatisfactory equals the probability the bag is overweight plus the probability it is underweight. That is, $P(A \text{ or } C) = P(A) + P(C) = .025 + .075 = .100$. The bag is satisfactory if it is not underweight or overweight, so $P(B) = 1 - [P(A) + P(C)] = 1 - [.025 + .075] = 0.900$. The Venn diagram portraying this situation is:



SELF-REVIEW 5–3

A sample of employees of Worldwide Enterprises is to be surveyed about a new health care plan. The employees are classified as follows:

Classification	Event	Number of Employees
Supervisors	*A*	120
Maintenance	*B*	50
Production	*C*	1,460
Management	*D*	302
Secretarial	*E*	68

(a) What is the probability that the first person selected is:
  (i) either in maintenance or a secretary?
  (ii) not in management?
(b) Draw a Venn diagram illustrating your answers to part (a).
(c) Are the events in part (a)(i) complementary or mutually exclusive or both?

The General Rule of Addition

The outcomes of an experiment may not be mutually exclusive. For example, the Florida Tourist Commission selected a sample of 200 tourists who visited the state during the year. The survey revealed that 120 tourists went to Disney World and 100 went to Busch Gardens near Tampa. What is the probability that a person selected visited either Disney World or Busch Gardens? If the special rule of addition is used, the probability of selecting a tourist who went to Disney World is .60, found by 120/200. Similarly, the probability of a tourist going to Busch Gardens is .50. The sum of these probabilities is 1.10. We know, however, that this probability cannot be greater than 1. The explanation is that many tourists visited both attractions and are being counted twice! A check of the survey responses revealed that 60 out of 200 sampled did, in fact, visit both attractions.

To answer our question, "What is the probability a selected person visited either Disney World or Busch Gardens?" (1) add the probability that a tourist visited Disney World and the probability he or she visited Busch Gardens, and (2) subtract the probability of visiting both. Thus:

$$P(\text{Disney or Busch}) = P(\text{Disney}) + P(\text{Busch}) - P(\text{both Disney and Busch})$$
$$= .60 + .50 - .30 = .80$$

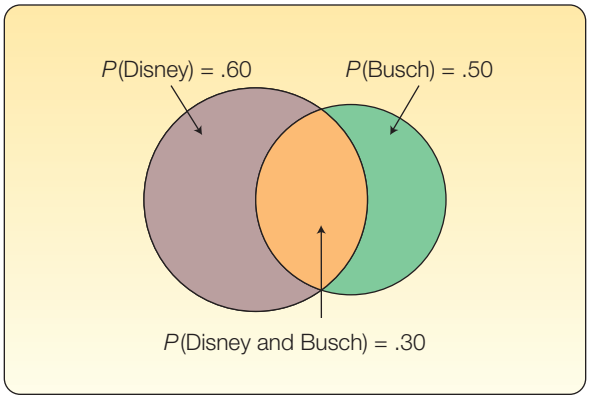
When two events both occur, the probability is called a **joint probability.** The probability (.30) that a tourist visits both attractions is an example of a joint probability.

© Rostislav Glinsky/Shutterstock.com

The following Venn diagram shows two events that are not mutually exclusive. The two events overlap to illustrate the joint event that some people have visited both attractions.

STATISTICS IN ACTION

If you wish to get some attention at the next gathering you attend, announce that you believe that at least two people present were born on the same date—that is, the same day of the year but not necessarily the same year. If there are 30 people in the room, the probability of a duplicate is .706. If there are 60 people in the room, the probability is .994 that at least two people share the same birthday. With as few as 23 people the chances are even, that is .50, that at least two people share the same birthday. Hint: To compute this, find the probability everyone was born on a different day and use the complement rule. Try this in your class.



JOINT PROBABILITY A probability that measures the likelihood two or more events will happen concurrently.

So the general rule of addition, which is used to compute the probability of two events that are not mutually exclusive, is:

GENERAL RULE OF ADDITION $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ **[5–4]**

For the expression $P(A \text{ or } B)$, the word *or* suggests that A may occur or B may occur. This also includes the possibility that A and B may occur. This use of *or* is sometimes called an **inclusive.** You could also write $P(A \text{ or } B \text{ or both})$ to emphasize that the union of the events includes the intersection of A and B.

If we compare the general and special rules of addition, the important difference is determining if the events are mutually exclusive. If the events *are* mutually exclusive, then the joint probability $P(A \text{ and } B)$ is 0 and we could use the special rule of addition. Otherwise, we must account for the joint probability and use the general rule of addition.

EXAMPLE

What is the probability that a card chosen at random from a standard deck of cards will be either a king or a heart?

SOLUTION

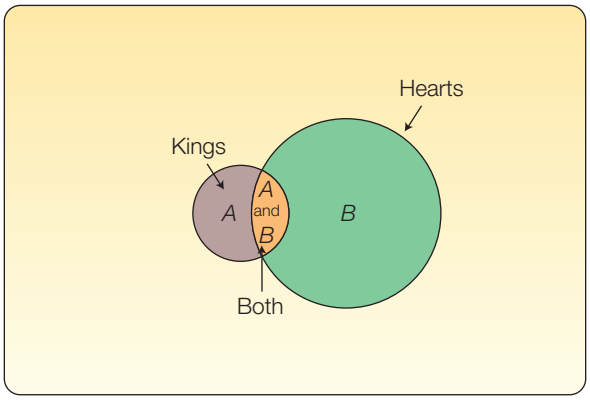
We may be inclined to add the probability of a king and the probability of a heart. But this creates a problem. If we do that, the king of hearts is counted with the kings and also with the hearts. So, if we simply add the probability of a king (there are 4 in a deck of 52 cards) to the probability of a heart (there are 13 in a deck of 52 cards) and report that 17 out of 52 cards meet the requirement, we have counted the king of hearts twice. We need to subtract 1 card from the 17 so the king of hearts is counted only once. Thus, there are 16 cards that are either hearts or kings. So the probability is 16/52 = .3077.

Card	Probability		Explanation
King	$P(A)$	= 4/52	4 kings in a deck of 52 cards
Heart	$P(B)$	= 13/52	13 hearts in a deck of 52 cards
King of Hearts	$P(A \text{ and } B)$	= 1/52	1 king of hearts in a deck of 52 cards

From formula (5–4):

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
$$= 4/52 + 13/52 - 1/52$$
$$= 16/52, \text{ or } .3077$$

A Venn diagram portrays these outcomes, which are not mutually exclusive.



SELF-REVIEW 5–4

Routine physical examinations are conducted annually as part of a health service program for General Concrete Inc. employees. It was discovered that 8% of the employees need corrective shoes, 15% need major dental work, and 3% need both corrective shoes and major dental work.

(a) What is the probability that an employee selected at random will need either corrective shoes or major dental work?
(b) Show this situation in the form of a Venn diagram.

EXERCISES

11. The events A and B are mutually exclusive. Suppose $P(A) = .30$ and $P(B) = .20$. What is the probability of either A or B occurring? What is the probability that neither A nor B will happen?

12. The events X and Y are mutually exclusive. Suppose $P(X) = .05$ and $P(Y) = .02$. What is the probability of either X or Y occurring? What is the probability that neither X nor Y will happen?

13. FILE A study of 200 advertising firms revealed their income after taxes:

Income after Taxes	Number of Firms
Under $1 million	102
$1 million to $20 million	61
$20 million or more	37

a. What is the probability an advertising firm selected at random has under $1 million in income after taxes?
b. What is the probability an advertising firm selected at random has either an income between $1 million and $20 million, or an income of $20 million or more? What rule of probability was applied?

14. The chair of the board of directors says, "There is a 50% chance this company will earn a profit, a 30% chance it will break even, and a 20% chance it will lose money next quarter."
a. Use an addition rule to find the probability the company will not lose money next quarter.
b. Use the complement rule to find the probability it will not lose money next quarter.

15. Suppose the probability you will get an A in this class is .25 and the probability you will get a B is .50. What is the probability your grade will be above a C?

16. Two coins are tossed. If *A* is the event "two heads" and *B* is the event "two tails," are *A* and *B* mutually exclusive? Are they complements?
17. The probabilities of the events *A* and *B* are .20 and .30, respectively. The probability that both *A* and *B* occur is .15. What is the probability of either *A* or *B* occurring?
18. Let $P(X) = .55$ and $P(Y) = .35$. Assume the probability that they both occur is .20. What is the probability of either *X* or *Y* occurring?
19. Suppose the two events *A* and *B* are mutually exclusive. What is the probability of their joint occurrence?
20. A student is taking two courses, history and math. The probability the student will pass the history course is .60, and the probability of passing the math course is .70. The probability of passing both is .50. What is the probability of passing at least one?
21. The aquarium at Sea Critters Depot contains 140 fish. Eighty of these fish are green swordtails (44 female and 36 male) and 60 are orange swordtails (36 female and 24 males). A fish is randomly captured from the aquarium:
    a. What is the probability the selected fish is a green swordtail?
    b. What is the probability the selected fish is male?
    c. What is the probability the selected fish is a male green swordtail?
    d. What is the probability the selected fish is either a male or a green swordtail?
22. A National Park Service survey of visitors to the Rocky Mountain region revealed that 50% visit Yellowstone Park, 40% visit the Tetons, and 35% visit both.
    a. What is the probability a vacationer will visit at least one of these attractions?
    b. What is the probability .35 called?
    c. Are the events mutually exclusive? Explain.

LO5-4
Calculate probabilities using the rules of multiplication.

RULES OF MULTIPLICATION TO CALCULATE PROBABILITY

In this section, we discuss the rules for computing the likelihood that two events both happen, or their joint probability. For example, 16% of the 2016 tax returns were prepared by H&R Block and 75% of those returns showed a refund. What is the likelihood a person's tax form was prepared by H&R Block and the person received a refund? Venn diagrams illustrate this as the intersection of two events. To find the likelihood of two events happening, we use the rules of multiplication. There are two rules of multiplication: the special rule and the general rule.

Special Rule of Multiplication

The special rule of multiplication requires that two events *A* and *B* are **independent.** Two events are independent if the occurrence of one event does not alter the probability of the occurrence of the other event.

INDEPENDENCE The occurrence of one event has no effect on the probability of the occurrence of another event.

One way to think about independence is to assume that events *A* and *B* occur at different times. For example, when event *B* occurs after event *A* occurs, does *A* have any effect on the likelihood that event *B* occurs? If the answer is no, then *A* and *B* are independent events. To illustrate independence, suppose two coins are tossed. The outcome of a coin toss (head or tail) is unaffected by the outcome of any other prior coin toss (head or tail).

For two independent events *A* and *B*, the probability that *A* and *B* will both occur is found by multiplying the two probabilities. This is the **special rule of multiplication** and is written symbolically as:

SPECIAL RULE OF MULTIPLICATION $P(A \text{ and } B) = P(A)P(B)$ **[5–5]**

For three independent events, A, B, and C, the special rule of multiplication used to determine the probability that all three events will occur is:

$$P(A \text{ and } B \text{ and } C) = P(A)P(B)P(C)$$

STATISTICS IN ACTION

In 2000 George W. Bush won the U.S. presidency by the slimmest of margins. Many election stories resulted, some involving voting irregularities, others raising interesting election questions. In a local Michigan election, there was a tie between two candidates for an elected position. To break the tie, the candidates drew a slip of paper from a box that contained two slips of paper, one marked "Winner" and the other unmarked. To determine which candidate drew first, election officials flipped a coin. The winner of the coin flip also drew the winning slip of paper. But was the coin flip really necessary? No, because the two events are independent. Winning the coin flip did not alter the probability of either candidate drawing the winning slip of paper.

EXAMPLE

A survey by the American Automobile Association (AAA) revealed 60% of its members made airline reservations last year. Two members are selected at random. What is the probability both made airline reservations last year?

SOLUTION

The probability the first member made an airline reservation last year is .60, written $P(R_1) = .60$, where R_1 refers to the fact that the first member made a reservation. The probability that the second member selected made a reservation is also .60, so $P(R_2) = .60$. Because the number of AAA members is very large, you may assume that R_1 and R_2 are independent. Consequently, using formula (5–5), the probability they both make a reservation is .36, found by:

$$P(R_1 \text{ and } R_2) = P(R_1)P(R_2) = (.60)(.60) = .36$$

All possible outcomes can be shown as follows. R means a reservation is made, and $\sim R$ means no reservation is made.

With the probabilities and the complement rule, we can compute the joint probability of each outcome. For example, the probability that neither member makes a reservation is .16. Further, the probability of the first or the second member (special addition rule) making a reservation is .48 (.24 + .24). You can also observe that the outcomes are mutually exclusive and collectively exhaustive. Therefore, the probabilities sum to 1.00.

Outcomes	Joint Probability	
R_1 R_2	(.60)(.60) =	.36
R_1 $\sim R_2$	(.60)(.40) =	.24
$\sim R_1$ R_2	(.40)(.60) =	.24
$\sim R_1$ $\sim R_2$	(.40)(.40) =	.16
Total		1.00

SELF-REVIEW 5–5

From experience, Teton Tire knows the probability is .95 that a particular XB-70 tire will last 60,000 miles before it becomes bald or fails. An adjustment is made on any tire that does not last 60,000 miles. You purchase four XB-70s. What is the probability all four tires will last at least 60,000 miles?

General Rule of Multiplication

If two events are not independent, they are referred to as **dependent.** To illustrate dependency, suppose there are 10 cans of soda in a cooler; 7 are regular and 3 are diet. A can is selected from the cooler. The probability of selecting a can of diet soda is 3/10, and the probability of selecting a can of regular soda is 7/10. Then a second can is selected from the cooler, without returning the first. The probability the second is diet depends on whether the first one selected was diet or not. The probability that the second is diet is:

2/9, if the first can is diet. (Only two cans of diet soda remain in the cooler.)
3/9, if the first can selected is regular. (All three diet sodas are still in the cooler.)

The fraction 2/9 (or 3/9) is called a **conditional probability** because its value is conditional on (dependent on) whether a diet or regular soda was the first selection from the cooler.

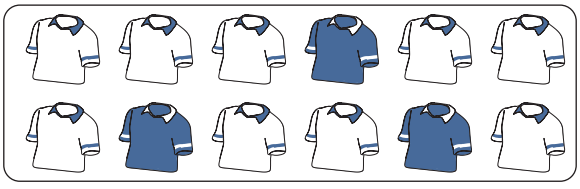
CONDITIONAL PROBABILITY The probability of a particular event occurring, given that another event has occurred.

In the general rule of multiplication, the conditional probability is required to compute the joint probability of two events that are not independent. For two events, A and B, that are not independent, the conditional probability is represented as $P(B|A)$, and expressed as the probability of B given A. Or the probability of B is conditional on the occurrence and effect of event A. Symbolically, the general rule of multiplication for two events that are not independent is:

GENERAL RULE OF MULTIPLICATION $P(A \text{ and } B) = P(A)P(B|A)$ **[5–6]**

EXAMPLE

A golfer has 12 golf shirts in his closet. Suppose 9 of these shirts are white and the others blue. He gets dressed in the dark, so he just grabs a shirt and puts it on. He plays golf two days in a row and does not launder and return the used shirts to the closet. What is the likelihood both shirts selected are white?



SOLUTION

The event that the first shirt selected is white is W_1. The probability is $P(W_1) = 9/12$ because 9 of the 12 shirts are white. The event that the second shirt selected is also white is identified as W_2. The conditional probability that the second shirt selected is white, given that the first shirt selected is also white, is $P(W_2|W_1) = 8/11$. Why is this so? Because after the first shirt is selected, there are only 11 shirts remaining in the closet and 8 of these are white. To determine the probability of 2 white shirts being selected, we use formula (5–6).

$$P(W_1 \text{ and } W_2) = P(W_1)P(W_2|W_1) = \left(\frac{9}{12}\right)\left(\frac{8}{11}\right) = .55$$

So the likelihood of selecting two shirts and finding them both to be white is .55.

We can extend the general rule of multiplication to more than two events. For three events A, B, and C, the formula is:

$$P(A \text{ and } B \text{ and } C) = P(A)P(B|A)P(C|A \text{ and } B)$$

In the case of the golf shirt example, the probability of selecting three white shirts without replacement is:

$$P(W_1 \text{ and } W_2 \text{ and } W_3) = P(W_1)P(W_2|W_1)P(W_3|W_1 \text{ and } W_2) = \left(\frac{9}{12}\right)\left(\frac{8}{11}\right)\left(\frac{7}{10}\right) = .38$$

So the likelihood of selecting three shirts without replacement and all being white is .38.

SELF-REVIEW 5–6

The board of directors of Tarbell Industries consists of eight men and four women. A four-member search committee is to be chosen at random to conduct a nationwide search for a new company president.

(a) What is the probability all four members of the search committee will be women?
(b) What is the probability all four members will be men?
(c) Does the sum of the probabilities for the events described in parts (a) and (b) equal 1? Explain.

LO5-5
Compute probabilities using a contingency table.

CONTINGENCY TABLES

Often we tally the results of a survey in a two-way table and use the results of this tally to determine various probabilities. We described this idea on page 116 in Chapter 4. To review, we refer to a two-way table as a **contingency table.**

CONTINGENCY TABLE A table used to classify sample observations according to two or more identifiable categories or classes.

A contingency table is a cross-tabulation that simultaneously summarizes two variables of interest and their relationship. The level of measurement can be nominal. Below are several examples.

- One hundred fifty adults were asked their gender and the number of Facebook accounts they used. The following table summarizes the results.

	Gender		
Facebook Accounts	**Men**	**Women**	**Total**
0	20	40	60
1	40	30	70
2 or more	10	10	20
Total	70	80	150

- The American Coffee Producers Association reports the following information on age and the amount of coffee consumed in a month.

	Coffee Consumption			
Age (Years)	**Low**	**Moderate**	**High**	**Total**
Under 30	36	32	24	92
30 up to 40	18	30	27	75
40 up to 50	10	24	20	54
50 and over	26	24	29	79
Total	90	110	100	300

According to this table, each of the 300 respondents is classified according to two criteria: (1) age and (2) the amount of coffee consumed.

The following example shows how the rules of addition and multiplication are used when we employ contingency tables.

EXAMPLE

Last month, the National Association of Theater Managers conducted a survey of 500 randomly selected adults. The survey asked respondents their age and the number of times they saw a movie in a theater. The results are summarized in Table 5–1.

TABLE 5–1 Number of Movies Attended per Month by Age

		Age			
Movies per Month		Less than 30 B_1	30 up to 60 B_2	60 or Older B_3	Total
0	A_1	15	50	10	75
1 or 2	A_2	25	100	75	200
3, 4, or 5	A_3	55	60	60	175
6 or more	A_4	5	15	30	50
Total		100	225	175	500

The association is interested in understanding the probabilities that an adult will see a movie in a theater, especially for adults 60 and older. This information is useful for making decisions regarding discounts on tickets and concessions for seniors.

Determine the probability of:

1. Selecting an adult who attended 6 or more movies per month.
2. Selecting an adult who attended 2 or fewer movies per month.
3. Selecting an adult who attended 6 or more movies per month **or** is 60 years of age or older.
4. Selecting an adult who attended 6 or more movies per month **given** the person is 60 years of age or older.
5. Selecting an adult who attended 6 or more movies per month **and** is 60 years of age or older.

Determine the independence of:

6. Number of movies per month attended and the age of the adult.

SOLUTION

Table 5–1 is called a contingency table. In a contingency table, an individual or an object is classified according to two criteria. In this example, a sampled adult is classified by age and by the number of movies attended per month. The rules of addition [formulas (5–2) and (5–4)] and the rules of multiplication [formulas (5–5) and (5–6)] allow us to answer the various probability questions based on the contingency table.

1. To find the probability that a randomly selected adult attended 6 or more movies per month, focus on the row labeled "6 or more" (also labeled A_4) in Table 5–1. The table shows that 50 of the total of 500 adults are in this class. Using the empirical approach, the probability is computed:

$$P(\text{6 or more}) = P(A_4) = \frac{50}{500} = .10$$

This probability indicates 10% of the 500 adults attend 6 or more movies per month.

2. To determine the probability of randomly selecting an adult who went to 2 or fewer movies per month, two outcomes must be combined: attending 0 movies per month and attending 1 or 2 movies per month. These two outcomes are mutually exclusive. That is, a person can only be classified as attending 0

movies per month, or 1 or 2 movies per month, not both. Because the two outcomes are mutually exclusive, we use the special rule of addition [formula (5–2)] by adding the probabilities of attending no movies and attending 1 or 2 movies:

$$P[(\text{attending 0}) \text{ or } (\text{attending 1 or 2})] = P(A_1) + P(A_2) = \left(\frac{75}{500} + \frac{200}{500}\right) = .55$$

So 55% of the adults in the sample attended 2 or fewer movies a month.

3. To determine the probability of randomly selecting an adult who went to "6 or more" movies per month or whose age is "60 or older," we again use the rules of addition. However, in this case the outcomes are **not** mutually exclusive. Why is this? Because a person can attend more than 6 movies per month, be 60 or older, or be both. So the two groups are not mutually exclusive because it is possible that a person would be counted in both groups. To determine this probability, the general rule of addition [formula (5–4)] is used.

$$P[(\text{6 or more}) \text{ or } (\text{60 or older})] = P(A_4) + P(B_3) - P(A_4 \text{ and } B_3)$$
$$= \left(\frac{50}{500} + \frac{175}{500} - \frac{30}{500}\right) = .39$$

So 39% of the adults are either 60 or older, attend 6 or more movies per month, or both.

4. To determine the probability of selecting a person who attends 6 or more movies per month given that the person is 60 or older, focus only on the column labeled B_3 in Table 5–1. That is, we are only interested in the 175 adults who are 60 or older. Of these 175 adults, 30 attended 6 or more movies. Using the general rule of multiplication [formula (5–6)]:

$$P[(\text{6 or more}) \text{ given } (\text{60 or older})] = P(A_4|B_3) = \frac{30}{175} = .17$$

Of the 500 adults, 17% of adults who are 60 or older attend 6 or more movies per month. This is called a conditional probability because the probability is based on the "condition" of being the age of 60 or older. Recall that in part (1), 10% of all adults attend 6 or more movies per month; here we see that 17% of adults who are 60 or older attend movies. This is valuable information for theater managers regarding the characteristics of their customers.

5. The probability a person attended 6 or more movies and is 60 or older is based on two conditions and they must both happen. That is, the two outcomes "6 or more movies" (A_4) and "60 or older" (B_3) must occur jointly. To find this joint probability we use the special rule of multiplication [formula (5–6)].

$$P[(\text{6 or more}) \text{ and } (\text{60 or older})] = P(A_4 \text{ and } B_3) = P(A_4)P(B_3|A_4)$$

To compute the joint probability, first compute the simple probability of the first outcome, A_4, randomly selecting a person who attends 6 or more movies. To find the probability, refer to row A_4 in Table 5–1. There are 50 of 500 adults that attended 6 or more movies. So $P(A_4) = 50/500$.

Next, compute the conditional probability $P(B_3|A_4)$. This is the probability of selecting an adult who is 60 or older given that the person attended 6 or more movies. The conditional probability is:

$$P[(\text{60 or older}) \text{ given } (\text{60 or more})] = P(B_3|A_4) = 30/50$$

Using these two probabilities, the joint probability that an adult attends 6 or more movies and is 60 or older is:

$$P[(\text{6 or more}) \text{ and } (\text{60 or older})] = P(A_4 \text{ and } B_3) = P(A_4)P(B_3|A_4)$$
$$= (50/500)(30/50) = .06$$

Based on the sample information from Table 5–1, the probability that an adult is both over 60 and attended 6 or more movies is 6%. It is important to know that the 6% is relative to all 500 adults.

Is there another way to determine this joint probability without using the special rule of multiplication formula? Yes. Look directly at the cell where row A_4, attends 6 or more movies, and column B_3, 60 or older, intersect. There are 30 adults in this cell that meet both criteria, so $P(A_4 \text{ and } B_3) = 30/500 = .06$. This is the same as computed with the formula.

6. Are the events independent? We can answer this question with the help of the results in part 4. In part 4 we found the probability of selecting an adult who was 60 or older given that the adult attended 6 or more movies was .17. If age is not a factor in movie attendance then we would expect the probability of a person who is 30 or less that attended 6 or more movies to also be 17%. That is, the two conditional probabilities would be the same. The probability that an adult attends 6 or more movies per month given the adult is less than 30 years old is:

$$P[(6 \text{ or more}) \text{ given } (\text{less than } 30)] = \frac{5}{100} = .05$$

Because these two probabilities are not the same, the number of movies attended and age are not independent. To put it another way, for the 500 adults, age is related to the number of movies attended. In Chapter 15, we investigate this concept of independence in greater detail.

SELF-REVIEW 5–7

Refer to Table 5–1 on page 151 to find the following probabilities.

(a) What is the probability of selecting an adult that is 30 up to 60 years old?
(b) What is the probability of selecting an adult who is under 60 years of age?
(c) What is the probability of selecting an adult who is less than 30 years old or attended no movies?
(d) What is the probability of selecting an adult who is less than 30 years old and went to no movies?

Tree Diagrams

A **tree diagram** is a visual that is helpful in organizing and calculating probabilities for problems similar to the previous example/solution. This type of problem involves several stages and each stage is illustrated with a branch of the tree. The branches of a tree diagram are labeled with probabilities. We will use the information in Table 5–1 to show the construction of a tree diagram.

1. We begin the construction by drawing a box with the variable, age, on the left to represent the root of the tree (see Chart 5–2).
2. There are three main branches going out from the root. The upper branch represents the outcome that an adult is less than 30 years old. The branch is labeled with the probability, $P(B_1) = 100/500$. The next branch represents the outcome that adults are 30 up to 60 years old. This branch is labeled with the probability $P(B_2) = 225/500$. The remaining branch is labeled $P(B_3) = 175/500$.
3. Four branches "grow" out of each of the four main branches. These branches represent the four categories of movies attended per month—0; 1 or 2; 3, 4, or 5; and 6 or more. The upper branches of the tree represent the conditional probabilities that an adult did not attend any movies given they are less than 30 years old. These are written $P(A_1|B_1)$, $P(A_2|B_1)$, $P(A_3|B_1)$, and $P(A_4|B_1)$ where A_1 refers to attending no movies; A_2 attending 1 or 2 movies per month; A_3 attending 3, 4, or 5 movies

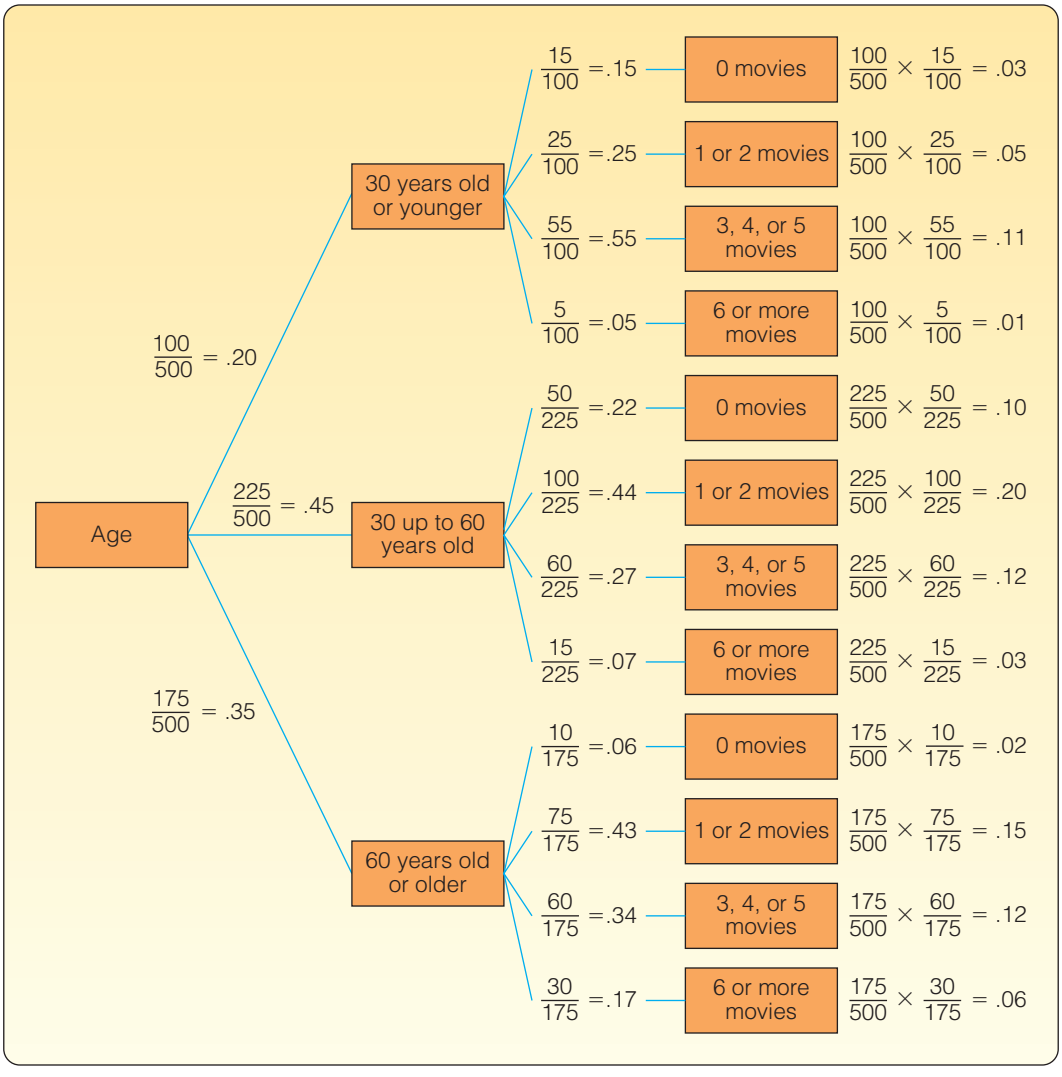


CHART 5–2 Tree Diagram Showing Age and Number of Movies Attended

per month; and A_4 attending 6 or more movies per month. For the upper branch of the tree, these probabilities are 15/100, 25/100, 55/100, and 5/100. We write the conditional probabilities in a similar fashion on the other branches.

4. Finally we determine the various joint probabilities. For the top branches, the events are an adult attends no movies per month and is 30 years old or younger; an adult attends 1 or 2 movies and is 30 years old or younger; an adult attends 3, 4, or 5 movies per month and is 30 years old or younger; and an adult attends 6 or more movies per month and is 30 years old or younger. These joint probabilities are shown on the right side of Chart 5–2. To explain, the joint probability that a randomly selected adult is less than 30 years old and attends 0 movies per month is:

$$P(B_1 \text{ and } A_1) = P(B_1)P(A_1|B_1) = \left(\frac{100}{500}\right)\left(\frac{15}{100}\right) = .03$$

The tree diagram summarizes all the probabilities based on the contingency table in Table 5–1. For example, the conditional probabilities show that the 60-and-older group has the highest percentage, 17%, attending 6 or movies per month. The

30-to-60-year-old group has the highest percentage, 22%, of seeing no movies per month. Based on the joint probabilities, 20% of the adults sampled attend 1 or 2 movies per month and are 30 up to 60 years of age. As you can see, there are many observations that we can make based on the information presented in the tree diagram.

SELF-REVIEW 5–8

Consumers were surveyed on the relative number of visits to a Sears store (often, occasional, and never) and if the store was located in an enclosed mall (yes and no). When variables are measured nominally, such as these data, the results are usually summarized in a contingency table.

	Enclosed Mall		
Visits	**Yes**	**No**	**Total**
Often	60	20	80
Occasional	25	35	60
Never	5	50	55
	90	105	195

What is the probability of selecting a shopper who:

(a) Visited a Sears store often?
(b) Visited a Sears store in an enclosed mall?
(c) Visited a Sears store in an enclosed mall or visited a Sears store often?
(d) Visited a Sears store often, given that the shopper went to a Sears store in an enclosed mall?

In addition:

(e) Are the number of visits and the enclosed mall variables independent?
(f) What is the probability of selecting a shopper who visited a Sears store often and it was in an enclosed mall?
(g) Draw a tree diagram and determine the various joint probabilities.

EXERCISES

23. Suppose $P(A) = .40$ and $P(B|A) = .30$. What is the joint probability of A and B?
24. Suppose $P(X_1) = .75$ and $P(Y_2|X_1) = .40$. What is the joint probability of X_1 and Y_2?
25. A local bank reports that 80% of its customers maintain a checking account, 60% have a savings account, and 50% have both. If a customer is chosen at random, what is the probability the customer has either a checking or a savings account? What is the probability the customer does not have either a checking or a savings account?
26. All Seasons Plumbing has two service trucks that frequently need repair. If the probability the first truck is available is .75, the probability the second truck is available is .50, and the probability that both trucks are available is .30, what is the probability neither truck is available?
27. FILE Refer to the following table.

	First Event			
Second Event	A_1	A_2	A_3	**Total**
B_1	2	1	3	6
B_2	1	2	1	4
Total	3	3	4	10

a. Determine $P(A_1)$.
b. Determine $P(B_1 | A_2)$.
c. Determine $P(B_2 \text{ and } A_3)$.

28. Three defective electric toothbrushes were accidentally shipped to a drugstore by Cleanbrush Products along with 17 nondefective ones.
a. What is the probability the first two electric toothbrushes sold will be returned to the drugstore because they are defective?
b. What is the probability the first two electric toothbrushes sold will not be defective?

29. FILE Each salesperson at Puchett, Sheets, and Hogan Insurance Agency is rated either below average, average, or above average with respect to sales ability. Each salesperson also is rated with respect to his or her potential for advancement—either fair, good, or excellent. These traits for the 500 salespeople were cross-classified into the following table.

	Potential for Advancement		
Sales Ability	**Fair**	**Good**	**Excellent**
Below average	16	12	22
Average	45	60	45
Above average	93	72	135

a. What is this table called?
b. What is the probability a salesperson selected at random will have above average sales ability and excellent potential for advancement?
c. Construct a tree diagram showing all the probabilities, conditional probabilities, and joint probabilities.

30. An investor owns three common stocks. Each stock, independent of the others, has equally likely chances of (1) increasing in value, (2) decreasing in value, or (3) remaining the same value. List the possible outcomes of this experiment. Estimate the probability at least two of the stocks increase in value.

31. FILE A survey of 545 college students asked: What is your favorite winter sport? And, what type of college do you attend? The results are summarized below:

	Favorite Winter Sport			
College Type	**Snowboarding**	**Skiing**	**Ice Skating**	**Total**
Junior College	68	41	46	155
Four-Year College	84	56	70	210
Graduate School	59	74	47	180
Total	211	171	163	545

Using these 545 students as the sample, a student from this study is randomly selected.
a. What is the probability of selecting a student whose favorite sport is skiing?
b. What is the probability of selecting a junior-college student?
c. If the student selected is a four-year-college student, what is the probability that the student prefers ice skating?
d. If the student selected prefers snowboarding, what is the probability that the student is in junior college?
e. If a graduate student is selected, what is the probability that the student prefers skiing or ice skating?

32. If you ask three strangers about their birthdays, what is the probability (a) All were born on Wednesday? (b) All were born on different days of the week? (c) None was born on Saturday?

LO5-6
Calculate probabilities using Bayes' theorem.

BAYES' THEOREM

In the 18th century, Reverend Thomas Bayes, an English Presbyterian minister, pondered this question: Does God really exist? Being interested in mathematics, he attempted to develop a formula to arrive at the probability God does exist based on evidence available to him on earth. Later Pierre-Simon Laplace refined Bayes' work and gave it the name "Bayes' theorem." The formula for **Bayes' theorem** is:

BAYES' THEOREM

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \qquad [5\text{–}7]$$

STATISTICS IN ACTION

A recent study by the National Collegiate Athletic Association (NCAA) reported that of 150,000 senior boys playing on their high school basketball team, 64 would make a professional team. To put it another way, the odds of a high school senior basketball player making a professional team are 1 in 2,344. From the same study:

1. The odds of a high school senior basketball player playing some college basketball are about 1 in 40.
2. The odds of a high school senior playing college basketball as a senior in college are about 1 in 60.
3. If you play basketball as a senior in college, the odds of making a professional team are about 1 in 37.5.

Assume in formula (5–7) that the events A_1 and A_2 are mutually exclusive and collectively exhaustive, and A_i refers to either event A_1 or A_2. Hence A_1 and A_2 are in this case complements. The meaning of the symbols used is illustrated by the following example.

Suppose 5% of the population of Umen, a fictional Third World country, have a disease that is peculiar to that country. We will let A_1 refer to the event "has the disease" and A_2 refer to the event "does not have the disease." Thus, we know that if we select a person from Umen at random, the probability the individual chosen has the disease is .05, or $P(A_1) = .05$. This probability, $P(A_1) = P(\text{has the disease}) = .05$, is called the **prior probability.** It is given this name because the probability is assigned before any empirical data are obtained.

PRIOR PROBABILITY The initial probability based on the present level of information.

The prior probability a person is not afflicted with the disease is therefore .95, or $P(A_2) = .95$, found by $1 - .05$.

There is a diagnostic technique to detect the disease, but it is not very accurate. Let B denote the event "test shows the disease is present." Assume that historical evidence shows that if a person actually has the disease, the probability that the test will indicate the presence of the disease is .90. Using the conditional probability definitions developed earlier in this chapter, this statement is written as:

$$P(B|A_1) = .90$$

Assume the probability is .15 that for a person who actually does not have the disease the test will indicate the disease is present.

$$P(B|A_2) = .15$$

Let's randomly select a person from Umen and perform the test. The test results indicate the disease is present. What is the probability the person actually has the disease? In symbolic form, we want to know $P(A_1|B)$, which is interpreted as: P(has the disease| the test results are positive). The probability $P(A_1|B)$ is called a **posterior probability.**

POSTERIOR PROBABILITY A revised probability based on additional information.

With the help of Bayes' theorem, formula (5–7), we can determine the posterior probability.

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$

$$= \frac{(.05)(.90)}{(.05)(.90) + (.95)(.15)} = \frac{.0450}{.1875} = .24$$

So the probability that a person has the disease, given that he or she tested positive, is .24. How is the result interpreted? If a person is selected at random from the population, the probability that he or she has the disease is .05. If the person is tested and the test result is positive, the probability that the person actually has the disease is increased about fivefold, from .05 to .24.

In the preceding problem, we had only two mutually exclusive and collectively exhaustive events, A_1 and A_2. If there are n such events, A_1, A_2, . . ., A_n, Bayes' theorem, formula (5–7), becomes

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)}$$

With the preceding notation, the calculations for the Umen problem are summarized in the following table.

Event, A_i	Prior Probability, $P(A_i)$	Conditional Probability, $P(B \mid A_i)$	Joint Probability, $P(A_i \text{ and } B)$	Posterior Probability, $P(A_i \mid B)$
Disease, A_1	.05	.90	.0450	.0450/.1875 = .24
No disease, A_2	.95	.15	.1425	.1425/.1875 = .76
			$P(B) = .1875$	1.00

Another illustration of Bayes' theorem follows.

EXAMPLE

© McGraw-Hill Education/ Marker Dierker, photographer

A manufacturer of cell phones purchases a microchip, called the LS-24, from three suppliers: Hall Electronics, Schuller Sales, and Crawford Components. Forty-five percent of the LS-24 chips are purchased from Hall Electronics, 30% from Schuller Sales, and the remaining 25% from Crawford Components. The manufacturer has extensive histories on the three suppliers and knows that 3% of the LS-24 chips from Hall Electronics are defective, 6% of chips from Schuller Sales are defective, and 4% of the chips purchased from Crawford Components are defective.

When the LS-24 chips arrive from the three suppliers, they are placed directly in a bin and not inspected or otherwise identified by supplier. A worker selects a chip for installation and finds it defective. What is the probability that it was manufactured by Schuller Sales?

SOLUTION

As a first step, let's summarize some of the information given in the problem statement.

- There are three mutually exclusive and collectively exhaustive events, that is, three suppliers.

  A_1 The LS-24 was purchased from Hall Electronics.
  A_2 The LS-24 was purchased from Schuller Sales.
  A_3 The LS-24 was purchased from Crawford Components.

- The prior probabilities are:

  $P(A_1) = .45$ The probability the LS-24 was manufactured by Hall Electronics.
  $P(A_2) = .30$ The probability the LS-24 was manufactured by Schuller Sales.
  $P(A_3) = .25$ The probability the LS-24 was manufactured by Crawford Components.

- The additional information can be either:

  B_1 The LS-24 is defective, or
  B_2 The LS-24 is not defective.

- The following conditional probabilities are given.

  $P(B_1|A_1) = .03$ The probability that an LS-24 chip produced by Hall Electronics is defective.
  $P(B_1|A_2) = .06$ The probability that an LS-24 chip produced by Schuller Sales is defective.
  $P(B_1|A_3) = .04$ The probability that an LS-24 chip produced by Crawford Components is defective.

- A chip is selected from the bin. Because the chips are not identified by supplier, we are not certain which supplier manufactured the chip. We want to determine the probability that the defective chip was purchased from Schuller Sales. The probability is written $P(A_2|B_1)$.

Look at Schuller's quality record. It is the worst of the three suppliers. They produce 30 percent of the product, but 6% are defective. Now that we have found a defective LS-24 chip, we suspect that $P(A_2|B_1)$ is greater than the 30% of $P(A_2)$. That is, we expect the revised probability to be greater than .30. But how much greater? Bayes' theorem can give us the answer. As a first step, consider the tree diagram in Chart 5–3.

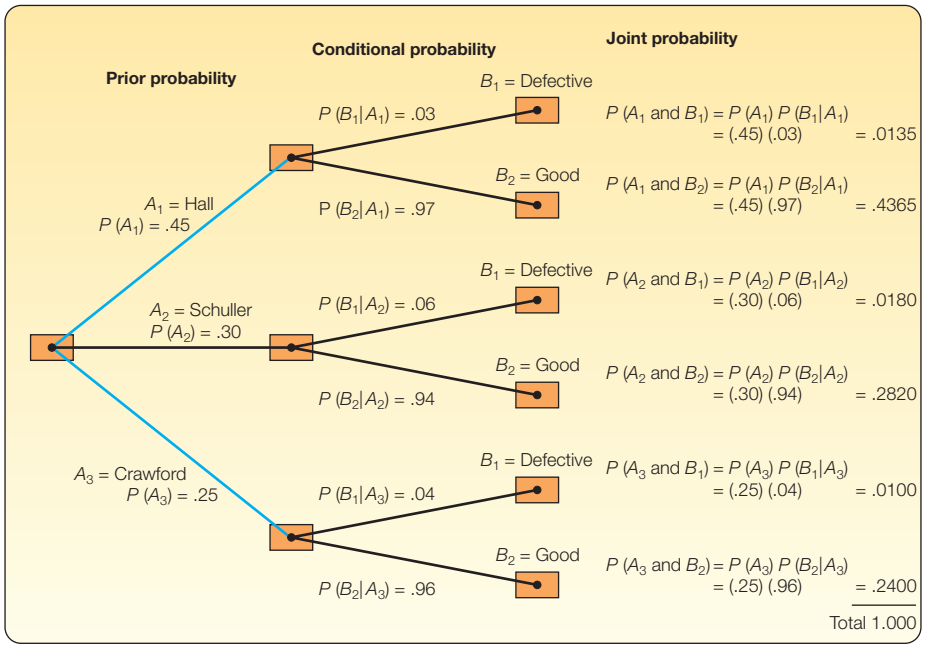

Prior probability
Conditional probability
Joint probability
A1 = Hall
P (A1) = .45
A2 = Schuller
P (A2) = .30
A3 = Crawford
P (A3) = .25
B1 = Defective
P (B1|A1) = .03
B2 = Good
P (B2|A1) = .97
B1 = Defective
P (B1|A2) = .06
B2 = Good
P (B2|A2) = .94
B1 = Defective
P (B1|A3) = .04
B2 = Good
P (B2|A3) = .96
P (A1 and B1) = P (A1) P (B1|A1) = (.45) (.03) = .0135
P (A1 and B2) = P (A1) P (B2|A1) = (.45) (.97) = .4365
P (A2 and B1) = P (A2) P (B1|A2) = (.30) (.06) = .0180
P (A2 and B2) = P (A2) P (B2|A2) = (.30) (.94) = .2820
P (A3 and B1) = P (A3) P (B1|A3) = (.25) (.04) = .0100
P (A3 and B2) = P (A3) P (B2|A3) = (.25) (.96) = .2400
Total 1.000

CHART 5–3 Tree Diagram of the Cell Phone Manufacturing Problem

REVISION OF CHART 5-3

The events are dependent, so the prior probability in the first branch is multiplied by the conditional probability in the second branch to obtain the joint probability. The joint probability is reported in the last column of Chart 5–3. To construct the tree diagram of Chart 5–3, we used a time sequence that moved from the supplier to the determination of whether the chip was defective.

What we need to do is reverse the time process. That is, instead of moving from left to right in Chart 5–3, we need to move from right to left. We have a defective chip, and we want to determine the likelihood that it was purchased from Schuller Sales. How is that accomplished? We first look at the joint probabilities as relative frequencies out of 10,000 cases. For example, the likelihood of a defective LS-24 chip that was produced by Hall Electronics is .0135. So of 10,000 cases, we would expect to find 135 defective chips produced by Hall Electronics. We observe that in 415 of 10,000 cases the LS-24 chip selected for assembly is defective, found by 135 + 180 + 100. Of these 415 defective chips, 180 were produced by Schuller Sales. Thus, the probability that the defective LS-24 chip was purchased from Schuller Sales is 180/415 = .4337. We have now determined the revised probability of $P(A_2|B_1)$. Before we found the defective chip, the likelihood that it was purchased from Schuller Sales was .30. This likelihood has been increased to .4337. What have we accomplished by using Bayes' Theorem? Once we found the defective part, we conclude that it is much more likely it is a product of Schuller Sales. The increase in the probability is rather dramatic moving from .30 to .4337.

This information is summarized in the following table.

Event, A_i	Prior Probability, $P(A_i)$	Conditional Probability, $P(B_1 \mid A_i)$	Joint Probability, $P(A_i \text{ and } B_1)$	Posterior Probability, $P(A_i \mid B_1)$
Hall	.45	.03	.0135	.0135/.0415 = .3235
Schuller	.30	.06	.0180	.0180/.0415 = .4337
Crawford	.25	.04	.0100	.0100/.0415 = .2410
			$P(B_1)$ = .0415	1.0000

The probability the defective LS-24 chip came from Schuller Sales can be formally found by using Bayes' theorem. We compute $P(A_2|B_1)$, where A_2 refers to Schuller Sales and B_1 to the fact that the selected LS-24 chip was defective.

$$P(A_2 \mid B_1) = \frac{P(A_2)P(B_1 \mid A_2)}{P(A_1)P(B_1 \mid A_1) + P(A_2)P(B_1 \mid A_2) + P(A_3)(B_1 \mid A_3)}$$

$$= \frac{(.30)(.06)}{(.45)(.03) + (.30)(.06) + (.25)(.04)} = \frac{.0180}{.04850} = .4337$$

This is the same result obtained from Chart 5–3 and from the conditional probability table.

SELF-REVIEW 5–9

Refer to the preceding example and solution.

(a) Design a formula to find the probability the part selected came from Crawford Components, given that it was a good chip.
(b) Compute the probability using Bayes' theorem.

EXERCISES

33. $P(A_1) = .60$, $P(A_2) = .40$, $P(B_1|A_1) = .05$, and $P(B_1|A_2) = .10$. Use Bayes' theorem to determine $P(A_1|B_1)$.

34. $P(A_1) = .20$, $P(A_2) = .40$, $P(A_3) = .40$, $P(B_1|A_1) = .25$, $P(B_1|A_2) = .05$, and $P(B_1|A_3) = .10$. Use Bayes' theorem to determine $P(A_3|B_1)$.

35. The Ludlow Wildcats baseball team, a minor league team in the Cleveland Indians organization, plays 70% of their games at night and 30% during the day. The team wins 50% of their night games and 90% of their day games. According to today's newspaper, they won yesterday. What is the probability the game was played at night?

36. Dr. Stallter has been teaching basic statistics for many years. She knows that 80% of the students will complete the assigned problems. She has also determined that among those who do their assignments, 90% will pass the course. Among those students who do not do their homework, 60% will pass. Mike Fishbaugh took statistics last semester from Dr. Stallter and received a passing grade. What is the probability that he completed the assignments?

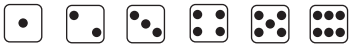
37. The credit department of Lion's Department Store in Anaheim, California, reported that 30% of their sales are cash, 30% are paid with a credit card, and 40% with a debit card. Twenty percent of the cash purchases, 90% of the credit card purchases, and 60% of the debit card purchases are for more than $50. Ms. Tina Stevens just purchased a new dress that cost $120. What is the probability that she paid cash?

38. One-fourth of the residents of the Burning Ridge Estates leave their garage doors open when they are away from home. The local chief of police estimates that 5% of the garages with open doors will have something stolen, but only 1% of those closed will have something stolen. If a garage is robbed, what is the probability the doors were left open?

LO5-7
Determine the number of outcomes using principles of counting.

PRINCIPLES OF COUNTING

If the number of possible outcomes in an experiment is small, it is relatively easy to count them. There are six possible outcomes, for example, resulting from the roll of a die, namely:



If, however, there are a large number of possible outcomes, such as the number of heads and tails for an experiment with 10 tosses, it would be tedious to count all the possibilities. They could have all heads, one head and nine tails, two heads and eight tails, and so on. To facilitate counting, we describe three formulas: the multiplication formula (not to be confused with the multiplication *rule* described earlier in the chapter), the permutation formula, and the combination formula.

The Multiplication Formula

We begin with the **multiplication formula.**

MULTIPLICATION FORMULA If there are m ways of doing one thing and n ways of doing another thing, there are $m \times n$ ways of doing both.

In terms of a formula:

MULTIPLICATION FORMULA Total number of arrangements $= (m)(n)$ **[5–8]**

This can be extended to more than two events. For three events m, n, and o:

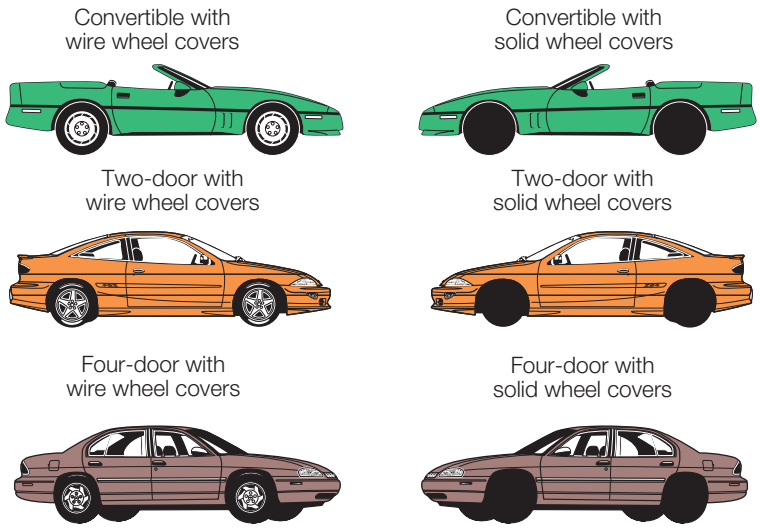
$$\text{Total number of arrangements} = (m)(n)(o)$$

EXAMPLE

An automobile dealer wants to advertise that for $29,999 you can buy a convertible, a two-door sedan, or a four-door model with your choice of either wire wheel covers or solid wheel covers. Based on the number of models and wheel covers, how many different vehicles can the dealer offer?

SOLUTION

Of course, the dealer could determine the total number of different cars by picturing and counting them. There are six.


Convertible with wire wheel covers
Convertible with solid wheel covers
Two-door with wire wheel covers
Two-door with solid wheel covers
Four-door with wire wheel covers
Four-door with solid wheel covers

We can employ the multiplication formula as a check (where m is the number of models and n the wheel cover type). From formula (5–8):

$$\text{Total possible arrangements} = (m)(n) = (3)(2) = 6$$

It was not difficult to count all the possible model and wheel cover combinations in this example. Suppose, however, that the dealer decided to offer eight models and six types of wheel covers. It would be tedious to picture and count all the possible alternatives. Instead, the multiplication formula can be used. In this case, there are $(m)(n) = (8)(6) = 48$ possible arrangements.

Note in the preceding applications of the multiplication formula that there were *two or more groupings from which you made selections*. The automobile dealer, for example, offered a choice of models and a choice of wheel covers. If a home builder offered you four different exterior styles of a home to choose from and three interior floor plans, the multiplication formula would be used to find how many different arrangements were possible. There are 12 possibilities.

SELF-REVIEW 5–10

1. The Women's Shopping Network on cable TV offers sweaters and slacks for women. The sweaters and slacks are offered in coordinating colors. If sweaters are available in five colors and the slacks are available in four colors, how many different outfits can be advertised?

2. Pioneer manufactures three models of Wifi Internet radios, two MP3 docking stations, four different sets of speakers, and three CD carousel changers. When the four types of components are sold together, they form a "system." How many different systems can the electronics firm offer?

The Permutation Formula

The multiplication formula is applied to find the number of possible arrangements for two or more groups. In contrast, we use the **permutation formula** to find the number of possible arrangements when there is a single group of objects. Illustrations of this type of problem are:

- Three electronic parts, a transistor, an LED, and a synthesizer, are assembled into a plug-in component for a HDTV. The parts can be assembled in any order. How many different ways can the three parts be assembled?
- A machine operator must make four safety checks before starting his machine. It does not matter in which order the checks are made. In how many different ways can the operator make the checks?

One order for the first illustration might be the transistor first, the LED second, and the synthesizer third. This arrangement is called a **permutation.**

PERMUTATION Any arrangement of r objects selected from a single group of n possible objects.

Note that the arrangements *a b c* and *b a c* are different permutations. The formula to count the total number of different permutations is:

PERMUTATION FORMULA

$$_nP_r = \frac{n!}{(n-r)!} \qquad [5\text{–}9]$$

where:

n is the total number of objects.
r is the number of objects selected.

Before we solve the two problems illustrated, the permutations and combinations (to be discussed shortly) use a notation called *n factorial*. It is written $n!$ and means the product of $n\,(n-1)(n-2)(n-3)\cdots(1)$. For instance, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Many of your calculators have a button with $x!$ that will perform this calculation for you. It will save you a great deal of time. For example the Texas Instrument Pro Scientific calculator has the following key:

It is the "third function," so check your users' manual or the Internet for instructions.

The factorial notation can also be canceled when the same number appears in both the numerator and the denominator, as shown below.

$$\frac{6!3!}{4!} = \frac{6 \cdot 5 \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}(3 \cdot 2 \cdot 1)}{\cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}} = 180$$

By definition, zero factorial, written 0!, is 1. That is, $0! = 1$.

EXAMPLE

Referring to the group of three electronic parts that are to be assembled in any order, in how many different ways can they be assembled?

SOLUTION

There are three electronic parts to be assembled, so $n = 3$. Because all three are to be inserted into the plug-in component, $r = 3$. Solving using formula (5–9) gives:

$$ {}_nP_r = \frac{n!}{(n-r)!} = \frac{3!}{(3-3)!} = \frac{3!}{0!} = \frac{3!}{1} = 6 $$

We can check the number of permutations arrived at by using the permutation formula. We determine how many "spaces" have to be filled and the possibilities for each "space." In the problem involving three electronic parts, there are three locations in the plug-in unit for the three parts. There are three possibilities for the first place, two for the second (one has been used up), and one for the third, as follows:

$$ (3)(2)(1) = 6 \text{ permutations} $$

The six ways in which the three electronic parts, lettered *A*, *B*, *C*, can be arranged are:

ABC *BAC* *CAB* *ACB* *BCA* *CBA*

In the previous example, we selected and arranged all the objects, that is $n = r$. In many cases, only some objects are selected and arranged from the n possible objects. We explain the details of this application in the following example.

EXAMPLE

The Fast Media Company is producing a one-minute video advertisement. In the production process, eight different video segments were made. To make the one-minute ad, they can only select three of the eight segments. How many different ways can the eight video segments be arranged in the three spaces available in the ad?

SOLUTION

There are eight possibilities for the first available space in the ad, seven for the second space (one has been used up), and six for the third space. Thus:

$$ (8)(7)(6) = 336, $$

that is, there are a total of 336 different possible arrangements. This could also be found by using formula (5–9). If $n = 8$ video segments and $r = 3$ spaces available, the formula leads to

$$ {}_nP_r = \frac{n!}{(n-r)!} = \frac{8!}{(8-3)!} = \frac{8!}{5!} = \frac{(8)(7)(6)\cancel{5!}}{\cancel{5!}} = 336 $$

The Combination Formula

If the order of the selected objects is *not* important, any selection is called a **combination.** Logically, the number of combinations is always less than the number of permutations. The formula to count the number of r object combinations from a set of n objects is:

COMBINATION FORMULA $$ {}_nC_r = \frac{n!}{r!(n-r)!} $$ **[5–10]**

For example, if executives Able, Baker, and Chauncy are to be chosen as a committee to negotiate a merger, there is only one possible combination of these three; the committee of Able, Baker, and Chauncy is the same as the committee of Baker, Chauncy, and Able. Using the combination formula:

$$ {}_nC_r = \frac{n!}{r!(n-r)!} = \frac{3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1(1)} = 1 $$

EXAMPLE

The Grand 16 movie theater uses teams of three employees to work the concession stand each evening. There are seven employees available to work each evening. How many different teams can be scheduled to staff the concession stand?

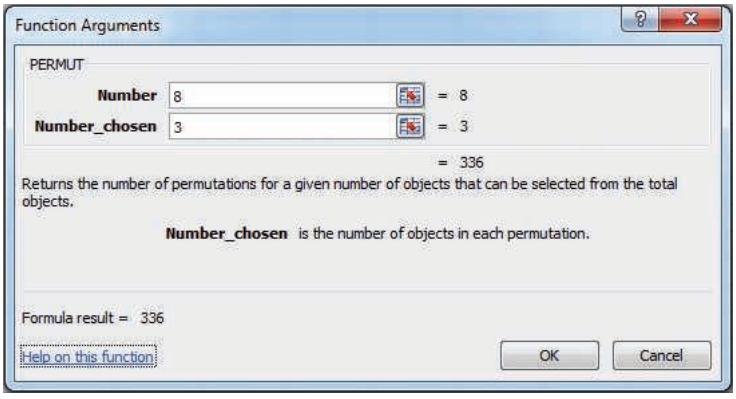
SOLUTION

According to formula (5–10), there are 35 combinations, found by

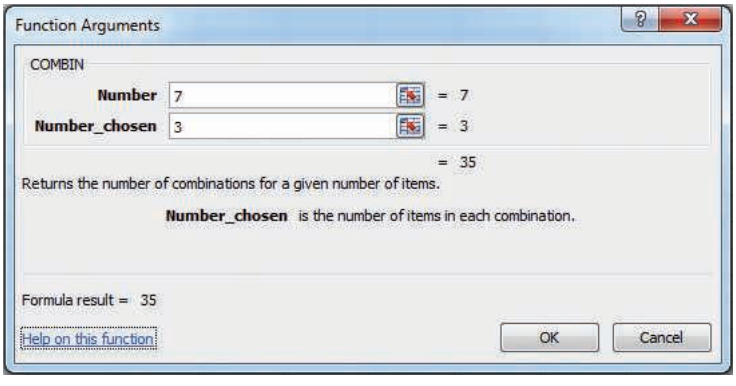
$$ {}_7C_3 = \frac{n!}{r!(n-r)!} = \frac{7!}{3!(7-3)!} = \frac{7!}{3!4!} = 35 $$

The seven employees taken three at a time would create the possibility of 35 different teams.

When the number of permutations or combinations is large, the calculations are tedious. Computer software and handheld calculators have "functions" to compute these numbers. The Excel output for the selection of three video segments for the eight available at the Fast Media Company is shown below. There are a total of 336 arrangements.


Function Arguments
PERMUT
Number 8 = 8
Number_chosen 3 = 3
= 336
Returns the number of permutations for a given number of objects that can be selected from the total objects.
Number_chosen is the number of objects in each permutation.
Formula result = 336
Help on this function
OK
Cancel

Below is the output for the number of teams at the Grand 16 movie theater. Three employees are chosen from seven possible employees.



SELF-REVIEW 5–11

1. A musician wants to write a score based on only five chords: B-flat, C, D, E, and G. However, only three chords out of the five will be used in succession, such as C, B-flat, and E. Repetitions, such as B-flat, B-flat, and E, will not be permitted.
   (a) How many permutations of the five chords, taken three at a time, are possible?
   (b) Using formula (5–9), how many permutations are possible?
2. The 10 numbers 0 through 9 are to be used in code groups of four to identify an item of clothing. Code 1083 might identify a blue blouse, size medium; the code group 2031 might identify a pair of pants, size 18; and so on. Repetitions of numbers are not permitted. That is, the same number cannot be used twice (or more) in a total sequence. For example, 2256, 2562, or 5559 would not be permitted. How many different code groups can be designed?
3. In the preceding example/solution involving the Grand 16 movie theater, there were 35 possible teams of three taken from seven employees.
   (a) Use formula (5–10) to show this is true.
   (b) The manager of the theater wants to plan for staffing the concession stand with teams of five employees on the weekends to serve the larger crowds. From the seven employees, how many teams of five employees are possible?
4. In a lottery game, three numbers are randomly selected from a tumbler of balls numbered 1 through 50.
   (a) How many permutations are possible?
   (b) How many combinations are possible?

EXERCISES

39. Solve the following:
   a. 40!/35!
   b. ${}_7P_4$
   c. ${}_5C_2$

40. Solve the following:
   a. 20!/17!
   b. ${}_9P_3$
   c. ${}_7C_2$

41. A pollster randomly selected 4 of 10 available people. How many different groups of 4 are possible?

42. A telephone number consists of seven digits, the first three representing the exchange. How many different telephone numbers are possible within the 537 exchange?

43. An overnight express company must include five cities on its route. How many different routes are possible, assuming that it does not matter in which order the cities are included in the routing?

44. A representative of the Environmental Protection Agency (EPA) wants to select samples from 10 landfills. The director has 15 landfills from which she can collect samples. How many different samples are possible?

45. Sam Snead's restaurant in Conway, South Carolina, offers an early bird special from 4–6 pm each week day evening. If each patron selects a Starter Selection (4 options), an Entrée (8 options), and a Dessert (3 options), how many different meals are possible?

46. A company is creating three new divisions and seven managers are eligible to be appointed head of a division. How many different ways could the three new heads be appointed? Hint: Assume the division assignment makes a difference.

CHAPTER SUMMARY

I. A probability is a value between 0 and 1 inclusive that represents the likelihood a particular event will happen.
- **A.** An experiment is the observation of some activity or the act of taking some measurement.
- **B.** An outcome is a particular result of an experiment.
- **C.** An event is the collection of one or more outcomes of an experiment.

II. There are three definitions of probability.
- **A.** The classical definition applies when there are n equally likely outcomes to an experiment.
- **B.** The empirical definition occurs when the number of times an event happens is divided by the number of observations.
- **C.** A subjective probability is based on whatever information is available.

III. Two events are mutually exclusive if by virtue of one event happening the other cannot happen.

IV. Events are independent if the occurrence of one event does not affect the occurrence of another event.

V. The rules of addition refer to the probability that any of two or more events can occur.
- **A.** The special rule of addition is used when events are mutually exclusive.

$$P(A \text{ or } B) = P(A) + P(B) \quad \textbf{[5–2]}$$

- **B.** The general rule of addition is used when the events are not mutually exclusive.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad \textbf{[5–4]}$$

- **C.** The complement rule is used to determine the probability of an event happening by subtracting the probability of the event not happening from 1.

$$P(A) = 1 - P(\sim A) \quad \textbf{[5–3]}$$

VI. The rules of multiplication are applied when two or more events occur simultaneously.
- **A.** The special rule of multiplication refers to events that are independent.

$$P(A \text{ and } B) = P(A)P(B) \quad \textbf{[5–5]}$$

- **B.** The general rule of multiplication refers to events that are not independent.

$$P(A \text{ and } B) = P(A)P(B|A) \quad \textbf{[5–6]}$$

- **C.** A joint probability is the likelihood that two or more events will happen at the same time.
- **D.** A conditional probability is the likelihood that an event will happen, given that another event has already happened.
- **E.** Bayes' theorem is a method of revising a probability, given that additional information is obtained. For two mutually exclusive and collectively exhaustive events:

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \quad \textbf{[5–7]}$$

VII. There are three counting rules that are useful in determining the number of outcomes in an experiment.
- **A.** The multiplication rule states that if there are m ways one event can happen and n ways another event can happen, then there are mn ways the two events can happen.

$$\text{Number of arrangements} = (m)(n) \quad \textbf{[5–8]}$$

- **B.** A permutation is an arrangement in which the order of the objects selected from a specific pool of objects is important.

$${}_nP_r = \frac{n!}{(n-r)!} \quad \textbf{[5–9]}$$

- **C.** A combination is an arrangement where the order of the objects selected from a specific pool of objects is not important.

$${}_nC_r = \frac{n!}{r!(n-r)!} \quad \textbf{[5–10]}$$

PRONUNCIATION KEY

SYMBOL	MEANING	PRONUNCIATION
$P(A)$	Probability of *A*	*P* of *A*
$P(\sim A)$	Probability of not *A*	*P* of not *A*
$P(A \text{ and } B)$	Probability of *A* and *B*	*P* of *A* and *B*
$P(A \text{ or } B)$	Probability of *A* or *B*	*P* of *A* or *B*
$P(A \mid B)$	Probability of *A* given *B* has happened	*P* of *A* given *B*
${}_nP_r$	Permutation of *n* items selected *r* at a time	*Pnr*
${}_nC_r$	Combination of *n* items selected *r* at a time	*Cnr*

CHAPTER EXERCISES

47. The marketing research department at Pepsico plans to survey teenagers about a newly developed soft drink. Each will be asked to compare it with his or her favorite soft drink.
- **a.** What is the experiment?
- **b.** What is one possible event?

48. The number of times a particular event occurred in the past is divided by the number of occurrences. What is this approach to probability called?

49. The probability that the cause and the cure for all cancers will be discovered before the year 2020 is .20. What viewpoint of probability does this statement illustrate?

50. FILE Berdine's Chicken Factory has several stores in the Hilton Head, South Carolina, area. When interviewing applicants for server positions, the owner would like to include information on the amount of tip a server can expect to earn per check (or bill). A study of 500 recent checks indicated the server earned the following amounts in tips per 8-hour shift.

Amount of Tip	Number
\$0 up to \$ 20	200
20 up to 50	100
50 up to 100	75
100 up to 200	75
200 or more	50
Total	500

- **a.** What is the probability of a tip of \$200 or more?
- **b.** Are the categories "\$0 up to \$20," "\$20 up to \$50," and so on considered mutually exclusive?
- **c.** If the probabilities associated with each outcome were totaled, what would that total be?
- **d.** What is the probability of a tip of up to \$50?
- **e.** What is the probability of a tip of less than \$200?

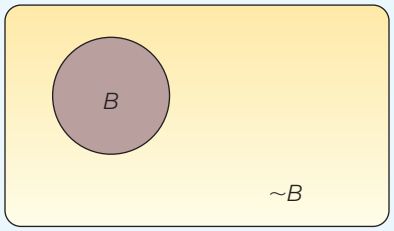
51. Winning all three "Triple Crown" races is considered the greatest feat of a pedigree racehorse. After a successful Kentucky Derby, Corn on the Cob is a heavy favorite at 2 to 1 odds to win the Preakness Stakes.
- **a.** If he is a 2 to 1 favorite to win the Belmont Stakes as well, what is his probability of winning the Triple Crown?
- **b.** What do his chances for the Preakness Stakes have to be in order for him to be "even money" to earn the Triple Crown?

52. The first card selected from a standard 52-card deck is a king.
- **a.** If it is returned to the deck, what is the probability that a king will be drawn on the second selection?
- **b.** If the king is not replaced, what is the probability that a king will be drawn on the second selection?

c. What is the probability that a king will be selected on the first draw from the deck and another king on the second draw (assuming that the first king was not replaced)?

53. Armco, a manufacturer of traffic light systems, found that under accelerated-life tests, 95% of the newly developed systems lasted 3 years before failing to change signals properly.
a. If a city purchased four of these systems, what is the probability all four systems would operate properly for at least 3 years?
b. Which rule of probability does this illustrate?
c. Using letters to represent the four systems, write an equation to show how you arrived at the answer to part (a).

54. Refer to the following picture.



a. What is the picture called?
b. What rule of probability is illustrated?
c. B represents the event of choosing a family that receives welfare payments. What does $P(B) + P(\sim B)$ equal?

55. In a management trainee program at Claremont Enterprises, 80% of the trainees are female and 20% male. Ninety percent of the females attended college, and 78% of the males attended college.
a. A management trainee is selected at random. What is the probability that the person selected is a female who did not attend college?
b. Are gender and attending college independent? Why?
c. Construct a tree diagram showing all the probabilities, conditional probabilities, and joint probabilities.
d. Do the joint probabilities total 1.00? Why?

56. Assume the likelihood that any flight on Delta Airlines arrives within 15 minutes of the scheduled time is .90. We randomly selected a Delta flight on four different days.
a. What is the likelihood all four of the selected flights arrived within 15 minutes of the scheduled time?
b. What is the likelihood that none of the selected flights arrived within 15 minutes of the scheduled time?
c. What is the likelihood at least one of the selected flights did not arrive within 15 minutes of the scheduled time?

57. There are 100 employees at Kiddie Carts International. Fifty-seven of the employees are hourly workers, 40 are supervisors, 2 are secretaries, and the remaining employee is the president. Suppose an employee is selected:
a. What is the probability the selected employee is an hourly worker?
b. What is the probability the selected employee is either an hourly worker or a supervisor?
c. Refer to part (b). Are these events mutually exclusive?
d. What is the probability the selected employee is neither an hourly worker nor a supervisor?

58. DJ LeMahieu of the Colorado Rockies had the highest batting average in the 2016 Major League Baseball season. His average was .348. So assume the probability of getting a hit is .348 for each time he batted. In a particular game, assume he batted three times.
a. This is an example of what type of probability?
b. What is the probability of getting three hits in a particular game?
c. What is the probability of not getting any hits in a game?
d. What is the probability of getting at least one hit?

59. Four women's college basketball teams are participating in a single-elimination holiday basketball tournament. If one team is favored in its semifinal match by odds of 2 to 1 and another squad is favored in its contest by odds of 3 to 1, what is the probability that:
a. Both favored teams win their games?
b. Neither favored team wins its game?
c. At least one of the favored teams wins its game?

60. There are three clues labeled "daily double" on the game show *Jeopardy.* If three equally matched contenders play, what is the probability that:
a. A single contestant finds all three "daily doubles"?
b. The returning champion gets all three of the "daily doubles"?
c. Each of the players selects precisely one of the "daily doubles"?

61. Brooks Insurance Inc. wishes to offer life insurance to men age 60 via the Internet. Mortality tables indicate the likelihood of a 60-year-old man surviving another year is .98. If the policy is offered to five men age 60:
a. What is the probability all five men survive the year?
b. What is the probability at least one does not survive?

62. Forty percent of the homes constructed in the Quail Creek area include a security system. Three homes are selected at random:
a. What is the probability all three of the selected homes have a security system?
b. What is the probability none of the three selected homes has a security system?
c. What is the probability at least one of the selected homes has a security system?
d. Did you assume the events to be dependent or independent?

63. Refer to Exercise 62, but assume there are 10 homes in the Quail Creek area and 4 of them have a security system. Three homes are selected at random:
a. What is the probability all three of the selected homes have a security system?
b. What is the probability none of the three selected homes has a security system?
c. What is the probability at least one of the selected homes has a security system?
d. Did you assume the events to be dependent or independent?

64. There are 20 families living in the Willbrook Farms Development. Of these families, 10 prepared their own federal income taxes for last year, 7 had their taxes prepared by a local professional, and the remaining 3 by H&R Block.
a. What is the probability of selecting a family that prepared their own taxes?
b. What is the probability of selecting two families, both of which prepared their own taxes?
c. What is the probability of selecting three families, all of which prepared their own taxes?
d. What is the probability of selecting two families, neither of which had their taxes prepared by H&R Block?

65. The board of directors of Saner Automatic Door Company consists of 12 members, 3 of whom are women. A new policy and procedures manual is to be written for the company. A committee of three is randomly selected from the board to do the writing.
a. What is the probability that all members of the committee are men?
b. What is the probability that at least one member of the committee is a woman?

66. FILE A recent survey reported in *BloombergBusinessweek* dealt with the salaries of CEOs at large corporations and whether company shareholders made money or lost money.

	CEO Paid More Than $1 Million	CEO Paid Less Than $1 Million	Total
Shareholders made money	2	11	13
Shareholders lost money	4	3	7
Total	6	14	20

If a company is randomly selected from the list of 20 studied, what is the probability:
a. The CEO made more than $1 million?
b. The CEO made more than $1 million or the shareholders lost money?
c. The CEO made more than $1 million given the shareholders lost money?
d. Of selecting two CEOs and finding they both made more than $1 million?

67. Althoff and Roll, an investment firm in Augusta, Georgia, advertises extensively in the *Augusta Morning Gazette,* the newspaper serving the region. The *Gazette* marketing

staff estimates that 60% of Althoff and Roll's potential market read the newspaper. It is further estimated that 85% of those who read the *Gazette* remember the Althoff and Roll advertisement.

a. What percent of the investment firm's potential market sees and remembers the advertisement?

b. What percent of the investment firm's potential market sees, but does not remember, the advertisement?

68. An Internet company located in Southern California has season tickets to the Los Angeles Lakers basketball games. The company president always invites one of the four vice presidents to attend games with him, and claims he selects the person to attend at random. One of the four vice presidents has not been invited to attend any of the last five Lakers home games. What is the likelihood this could be due to chance?

69. A computer-supply retailer purchased a batch of 1,000 CD-R disks and attempted to format them for a particular application. There were 857 perfect CDs, 112 CDs were usable but had bad sectors, and the remainder could not be used at all.

a. What is the probability a randomly chosen CD is not perfect?

b. If the disk is not perfect, what is the probability it cannot be used at all?

70. An investor purchased 100 shares of Fifth Third Bank stock and 100 shares of Santee Electric Cooperative stock. The probability the bank stock will appreciate over a year is .70. The probability the electric utility will increase over the same period is .60. Assume the two events are independent.

a. What is the probability both stocks appreciate during the period?

b. What is the probability the bank stock appreciates but the utility does not?

c. What is the probability at least one of the stocks appreciates?

71. Flashner Marketing Research Inc. specializes in providing assessments of the prospects for women's apparel shops in shopping malls. Al Flashner, president, reports that he assesses the prospects as good, fair, or poor. Records from previous assessments show that 60% of the time the prospects were rated as good, 30% of the time fair, and 10% of the time poor. Of those rated good, 80% made a profit the first year; of those rated fair, 60% made a profit the first year; and of those rated poor, 20% made a profit the first year. Connie's Apparel was one of Flashner's clients. Connie's Apparel made a profit last year. What is the probability that it was given an original rating of poor?

72. Two boxes of men's Old Navy shirts were received from the factory. Box 1 contained 25 mesh polo shirts and 15 Super-T shirts. Box 2 contained 30 mesh polo shirts and 10 Super-T shirts. One of the boxes was selected at random, and a shirt was chosen at random from that box to be inspected. The shirt was a mesh polo shirt. Given this information, what is the probability that the mesh polo shirt came from Box 1?

73. With each purchase of a large pizza at Tony's Pizza, the customer receives a coupon that can be scratched to see if a prize will be awarded. The probability of winning a free soft drink is 0.10, and the probability of winning a free large pizza is 0.02. You plan to eat lunch tomorrow at Tony's. What is the probability:

a. That you will win either a large pizza or a soft drink?

b. That you will not win a prize?

c. That you will not win a prize on three consecutive visits to Tony's?

d. That you will win at least one prize on one of your next three visits to Tony's?

74. For the daily lottery game in Illinois, participants select three numbers between 0 and 9. A number cannot be selected more than once, so a winning ticket could be, say, 307 but not 337. Purchasing one ticket allows you to select one set of numbers. The winning numbers are announced on TV each night.

a. How many different outcomes (three-digit numbers) are possible?

b. If you purchase a ticket for the game tonight, what is the likelihood you will win?

c. Suppose you purchase three tickets for tonight's drawing and select a different number for each ticket. What is the probability that you will not win with any of the tickets?

75. Several years ago, Wendy's Hamburgers advertised that there are 256 different ways to order your hamburger. You may choose to have, or omit, any combination of the following on your hamburger: mustard, ketchup, onion, pickle, tomato, relish, mayonnaise, and lettuce. Is the advertisement correct? Show how you arrive at your answer.

76. Recent surveys indicate 60% of tourists to China visited the Forbidden City, the Temple of Heaven, the Great Wall, and other historical sites in or near Beijing. Forty percent visited Xi'an with its magnificent terra-cotta soldiers, horses, and chariots, which lay buried for over 2,000 years. Thirty percent of the tourists went to both Beijing and Xi'an. What is the probability that a tourist visited at least one of these places?

77. A new chewing gum has been developed that is helpful to those who want to stop smoking. If 60% of those people chewing the gum are successful in stopping smoking, what is the probability that in a group of four smokers using the gum at least one quits smoking?

78. Reynolds Construction Company has agreed not to erect all "look-alike" homes in a new subdivision. Five exterior designs are offered to potential home buyers. The builder has standardized three interior plans that can be incorporated in any of the five exteriors. How many different ways can the exterior and interior plans be offered to potential home buyers?

79. A new sports car model has defective brakes 15% of the time and a defective steering mechanism 5% of the time. Let's assume (and hope) that these problems occur independently. If one or the other of these problems is present, the car is called a "lemon." If both of these problems are present, the car is a "hazard." Your instructor purchased one of these cars yesterday. What is the probability it is:
   a. A lemon?
   b. A hazard?

80. The state of Maryland has license plates with three numbers followed by three letters. How many different license plates are possible?

81. There are four people being considered for the position of chief executive officer of Dalton Enterprises. Three of the applicants are over 60 years of age. Two are female, of which only one is over 60.
   a. What is the probability that a candidate is over 60 and female?
   b. Given that the candidate is male, what is the probability he is less than 60?
   c. Given that the person is over 60, what is the probability the person is female?

82. Tim Bleckie is the owner of Bleckie Investment and Real Estate Company. The company recently purchased four tracts of land in Holly Farms Estates and six tracts in Newburg Woods. The tracts are all equally desirable and sell for about the same amount.
   a. What is the probability that the next two tracts sold will be in Newburg Woods?
   b. What is the probability that of the next four sold at least one will be in Holly Farms?
   c. Are these events independent or dependent?

83. A computer password consists of four characters. The characters can be one of the 26 letters of the alphabet. Each character may be used more than once. How many different passwords are possible?

84. A case of 24 cans contains 1 can that is contaminated. Three cans are to be chosen randomly for testing.
   a. How many different combinations of three cans could be selected?
   b. What is the probability that the contaminated can is selected for testing?

85. A puzzle in the newspaper presents a matching problem. The names of 10 U.S. presidents are listed in one column, and their vice presidents are listed in random order in the second column. The puzzle asks the reader to match each president with his vice president. If you make the matches randomly, how many matches are possible? What is the probability all 10 of your matches are correct?

86. Two components, *A* and *B*, operate in series. Being in series means that for the system to operate, both components *A* and *B* must work. Assume the two components are independent. What is the probability the system works under these conditions? The probability *A* works is .90 and the probability *B* functions is also .90.

87. Horwege Electronics Inc. purchases TV picture tubes from four different suppliers. Tyson Wholesale supplies 20% of the tubes, Fuji Importers 30%, Kirkpatricks 25%, and Parts Inc. 25%. Tyson Wholesale tends to have the best quality, as only 3% of its tubes arrive defective. Fuji Importers' tubes are 4% defective, Kirkpatricks' 7%, and Parts Inc.'s are 6.5% defective.
   a. What is the overall percent defective?

b. A defective picture tube was discovered in the latest shipment. What is the probability that it came from Tyson Wholesale?

88. ABC Auto Insurance classifies drivers as good, medium, or poor risks. Drivers who apply to them for insurance fall into these three groups in the proportions 30%, 50%, and 20%, respectively. The probability a "good" driver will have an accident is .01, the probability a "medium" risk driver will have an accident is .03, and the probability a "poor" driver will have an accident is .10. The company sells Mr. Brophy an insurance policy and he has an accident. What is the probability Mr. Brophy is:
a. A "good" driver?
b. A "medium" risk driver?
c. A "poor" driver?

89. You take a trip by air that involves three independent flights. If there is an 80% chance each specific leg of the trip is on time, what is the probability all three flights arrive on time?

90. The probability a D-Link network server is down is .05. If you have three independent servers, what is the probability that at least one of them is operational?

91. Twenty-two percent of all light emitting diode (LED) displays are manufactured by Samsung. What is the probability that in a collection of three independent LED HDTV purchases, at least one is a Samsung?

DATA ANALYTICS

92. FILE Refer to the North Valley Real Estate data, which report information on homes sold during the last year.
a. Sort the data into a table that shows the number of homes that have a pool versus the number that don't have a pool in each of the five townships. If a home is selected at random, compute the following probabilities.
1. The home has a pool.
2. The home is in Township 1 or has a pool.
3. Given that it is in Township 3, that it has a pool.
4. The home has a pool and is in Township 3.
b. Sort the data into a table that shows the number of homes that have a garage attached versus those that don't in each of the five townships. If a home is selected at random, compute the following probabilities:
1. The home has a garage attached.
2. The home does not have a garage attached, given that it is in Township 5.
3. The home has a garage attached and is in Township 3.
4. The home does not have a garage attached or is in Township 2.

93. FILE Refer to the Baseball 2016 data, which reports information on the 30 Major League Baseball teams for the 2016 season. Set up three variables:
- Divide the teams into two groups, those that had a winning season and those that did not. That is, create a variable to count the teams that won 81 games or more, and those that won 80 or less.
- Create a new variable for attendance, using three categories: attendance less than 2.0 million, attendance of 2.0 million up to 3.0 million, and attendance of 3.0 million or more.
- Create a variable that shows the teams that play in a stadium less than 20 years old versus one that is 20 years old or more.

Answer the following questions.
a. Create a table that shows the number of teams with a winning season versus those with a losing season by the three categories of attendance. If a team is selected at random, compute the following probabilities:
1. The team had a winning season.
2. The team had a winning season or attendance of more than 3.0 million.
3. The team had a winning season given attendance was more than 3.0 million.
4. The team has a winning season and attracted fewer than 2.0 million fans.

b. Create a table that shows the number of teams with a winning season versus those that play in new or old stadiums. If a team is selected at random, compute the following probabilities:
   1. Selecting a team with a stadium that is at least 20 years old.
   2. The likelihood of selecting a team with a winning record and playing in a new stadium.
   3. The team had a winning record or played in a new stadium.

94. FILE Refer to the Lincolnville school bus data. Set up a variable that divides the age of the buses into three groups: new (less than 5 years old), medium (5 but less than 10 years), and old (10 or more years). The median maintenance cost is $4,179. Based on this value, create a variable for those less than or equal to the median (low maintenance) and those more than the median (high maintenance cost). Finally, develop a table to show the relationship between maintenance cost and age of the bus.
   a. What percentage of the buses are less than five years old?
   b. What percentage of the buses less than five years old have low maintenance costs?
   c. What percentage of the buses ten or more years old have high maintenance costs?
   d. Does maintenance cost seem to be related to the age of the bus? Hint: Compare the maintenance cost of the old buses with the cost of the new buses? Would you conclude maintenance cost is independent of the age?

13 Correlation and Linear Regression

© Ingram Publishing/Superstock RF

▲ **TRAVELAIR.COM** samples domestic airline flights to explore the relationship between airfare and distance. The service would like to know if there is a correlation between airfare and flight distance. If there is a correlation, what percentage of the variation in airfare is accounted for by distance? How much does each additional mile add to the fare? (See Exercise 61 and **LO13-2**, **LO13-3**, and **LO13-5**.)

LEARNING OBJECTIVES

When you have completed this chapter, you will be able to:

LO13-1 Explain the purpose of correlation analysis.

LO13-2 Calculate a correlation coefficient to test and interpret the relationship between two variables.

LO13-3 Apply regression analysis to estimate the linear relationship between two variables.

LO13-4 Evaluate the significance of the slope of the regression equation.

LO13-5 Evaluate a regression equation's ability to predict using the standard estimate of the error and the coefficient of determination.

LO13-6 Calculate and interpret confidence and prediction intervals.

LO13-7 Use a log function to transform a nonlinear relationship.

INTRODUCTION

Chapters 2 through 4 presented *descriptive statistics.* We organized raw data into a frequency distribution and computed several measures of location and measures of dispersion to describe the major characteristics of the distribution. In Chapters 5 through 7, we described probability, and from probability statements, we created probability distributions. In Chapters 8 through 12, we studied *statistical inference,* where we collected a sample to estimate a population parameter such as the population mean or population proportion. In addition, we used the sample data to test a hypothesis about a population mean or a population proportion, the difference between two population means, or the equality of several population means. Each of these tests involved just *one* interval- or ratio-level variable, such as the profit made on a car sale, the income of bank presidents, or the number of patients admitted each month to a particular hospital.

In this chapter, we shift the emphasis to the study of relationships between two interval- or ratio-level variables. In all business fields, identifying and studying relationships between variables can provide information on ways to increase profits, methods to decrease costs, or variables to predict demand. In marketing products, many firms use price reductions through coupons and discount pricing to increase sales. In this example, we are interested in the relationship between two variables: price reductions and sales. To collect the data, a company can test-market a variety of price reduction methods and observe sales. We hope to confirm a relationship that decreasing price leads to increased sales. In economics, you will find many relationships between two variables that are the basis of economics, such as price and demand.

As another familiar example, recall in Chapter 4 we used the Applewood Auto Group data to show the relationship between two variables with a scatter diagram. We plotted the profit for each vehicle sold on the vertical axis and the age of the buyer on the horizontal axis. See page 116. In that graph, we observed that as the age of the buyer increased, the profit for each vehicle also increased.

Other examples of relationships between two variables are:

- Does the amount Healthtex spends per month on training its sales force affect its monthly sales?
- Is the number of square feet in a home related to the cost to heat the home in January?
- In a study of fuel efficiency, is there a relationship between miles per gallon and the weight of a car?
- Does the number of hours that students study for an exam influence the exam score?

In this chapter, we carry this idea further. That is, we develop numerical measures to express the relationship between two variables. Is the relationship strong or weak? Is it direct or inverse? In addition, we develop an equation to express the relationship between variables. This will allow us to estimate one variable on the basis of another.

To begin our study of relationships between two variables, we examine the meaning and purpose of **correlation analysis.** We continue by developing an equation that will allow us to estimate the value of one variable based on the value of another. This is called **regression analysis.** We will also evaluate the ability of the equation to accurately make estimations.

STATISTICS IN ACTION

The space shuttle *Challenger* exploded on January 28, 1986. An investigation of the cause examined four contractors: Rockwell International for the shuttle and engines, Lockheed Martin for ground support, Martin Marietta for the external fuel tanks, and Morton Thiokol for the solid fuel booster rockets. After several months, the investigation blamed the explosion on defective O-rings produced by Morton Thiokol. A study of the contractor's stock prices showed an interesting happenstance. On the day of the *Challenger* explosion, Morton Thiokol stock was down 11.86% and the stock of the other three lost only 2 to 3%. Can we conclude that financial markets predicted the outcome of the investigation?

LO13-1
Explain the purpose of correlation analysis.

WHAT IS CORRELATION ANALYSIS?

When we study the relationship between two interval- or ratio-scale variables, we often start with a scatter diagram. This procedure provides a visual representation of the relationship between the variables. The next step is usually to calculate the correlation coefficient. It provides a quantitative measure of the strength of the relationship between

TABLE 13–1 Number of Sales Calls and Copiers Sold for 15 Salespeople

Sales Representative	Sales Calls	Copiers Sold
Brian Virost	96	41
Carlos Ramirez	40	41
Carol Saia	104	51
Greg Fish	128	60
Jeff Hall	164	61
Mark Reynolds	76	29
Meryl Rumsey	72	39
Mike Kiel	80	50
Ray Snarsky	36	28
Rich Niles	84	43
Ron Broderick	180	70
Sal Spina	132	56
Soni Jones	120	45
Susan Welch	44	31
Tom Keller	84	30

two variables. As an example, the sales manager of North American Copier Sales, which has a large sales force throughout the United States and Canada, wants to determine whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 15 representatives and determines, for each representative, the number of sales calls made and the number of copiers sold. This information is reported in Table 13–1.

By reviewing the data, we observe that there does seem to be some relationship between the number of sales calls and the number of units sold. That is, the salespeople who made the most sales calls sold the most units. However, the relationship is not "perfect" or exact. For example, Soni Jones made fewer sales calls than Jeff Hall, but she sold more units.

In addition to the graphical techniques in Chapter 4, we will develop numerical measures to precisely describe the relationship between the two variables, sales calls and copiers sold. This group of statistical techniques is called **correlation analysis.**

CORRELATION ANALYSIS A group of techniques to measure the relationship between two variables.

The basic idea of correlation analysis is to report the relationship between two variables. The usual first step is to plot the data in a **scatter diagram.** An example will show how a scatter diagram is used.

EXAMPLE

North American Copier Sales sells copiers to businesses of all sizes throughout the United States and Canada. Ms. Marcy Bancer was recently promoted to the position of national sales manager. At the upcoming sales meeting, the sales representatives from all over the country will be in attendance. She would like to impress upon them the importance of making that extra sales call each day. She decides to gather some information on the relationship between the number of sales calls and the number of copiers sold. She selects a random sample of 15 sales representatives and determines the number of sales calls they made last month and the number of copiers they sold. The sample information is reported in Table 13–1. What

observations can you make about the relationship between the number of sales calls and the number of copiers sold? Develop a scatter diagram to display the information.

SOLUTION

Based on the information in Table 13–1, Ms. Bancer suspects there is a relationship between the number of sales calls made in a month and the number of copiers sold. Ron Broderick sold the most copiers last month and made 180 sales calls. On the other hand, Ray Snarsky, Carlos Ramirez, and Susan Welch made the fewest calls: 36, 40, and 44. They also had the lowest number of copiers sold among the sampled representatives.

The implication is that the number of copiers sold is related to the number of sales calls made. As the number of sales calls increases, it appears the number of copiers sold also increases. We refer to number of sales calls as the **independent variable** and number of copiers sold as the **dependent variable.**

The independent variable provides the basis for estimating or predicting the dependent variable. For example, we would like to predict the expected number of copiers sold if a salesperson makes 100 sales calls. In the randomly selected sample data, the independent variable—sales calls—is a random number.

The dependent variable is the variable that is being predicted or estimated. It can also be described as the result or outcome for a particular value of the independent variable. The dependent variable is random. That is, for a given value of the independent variable, there are many possible outcomes for the dependent variable.

It is common practice to scale the dependent variable (copiers sold) on the vertical or *Y*-axis and the independent variable (number of sales calls) on the horizontal or *X*-axis. To develop the scatter diagram of the North American Copier Sales information, we begin with the first sales representative, Brian Virost. Brian made 96 sales calls last month and sold 41 copiers, so $x = 96$ and $y = 41$. To plot this point, move along the horizontal axis to $x = 96$, then go vertically to $y = 41$ and place a dot at the intersection. This process is continued until all the paired data are plotted, as shown in Chart 13–1.

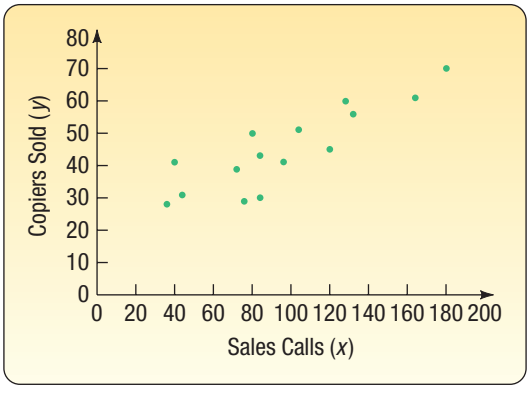


CHART 13–1 Scatter Diagram Showing Sales Calls and Copiers Sold

The scatter diagram shows graphically that the sales representatives who make more calls tend to sell more copiers. It is reasonable for Ms. Bancer, the national sales manager, to tell her salespeople that the more sales calls they make, the more copiers they can expect to sell. Note that, while there appears to be a positive relationship between the two variables, all the points do not fall on a straight line. In the following section, you will measure the strength and direction of this relationship between two variables by determining the correlation coefficient.

LO13-2
Calculate a correlation coefficient to test and interpret the relationship between two variables.

THE CORRELATION COEFFICIENT

Originated by Karl Pearson about 1900, the **correlation coefficient** describes the strength of the relationship between two sets of interval-scaled or ratio-scaled variables. Designated *r,* it is often referred to as *Pearson's r* and as the *Pearson product-moment correlation coefficient*. It can assume any value from −1.00 to +1.00 inclusive. A correlation coefficient of −1.00 or +1.00 indicates *perfect correlation*. For example, a correlation coefficient for the preceding example computed to be +1.00 would indicate that the number of sales calls and the number of copiers sold are perfectly related in a positive linear sense. A computed value of −1.00 would reveal that sales calls and the number of copiers sold are perfectly related in an inverse linear sense. How the scatter diagram would appear if the relationship between the two variables were linear and perfect is shown in Chart 13–2.

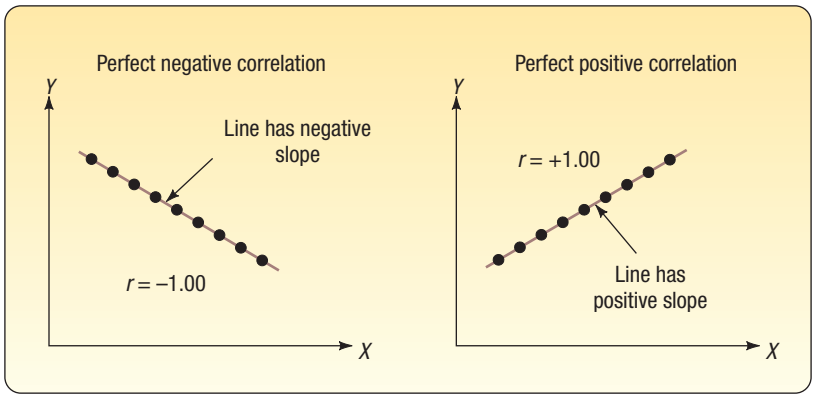


CHART 13–2 Scatter Diagrams Showing Perfect Negative Correlation and Perfect Positive Correlation

If there is absolutely no relationship between the two sets of variables, Pearson's *r* is zero. A correlation coefficient *r* close to 0 (say, .08) shows that the linear relationship is quite weak. The same conclusion is drawn if *r* = −.08. Coefficients of −.91 and +.91 have equal strength; both indicate very strong correlation between the two variables. Thus, *the strength of the correlation does not depend on the direction (either − or +).*

Scatter diagrams for *r* = 0, a weak *r* (say, −.23), and a strong *r* (say, +.87) are shown in Chart 13–3. Note that, if the correlation is weak, there is considerable scatter about a line drawn through the center of the data. For the scatter diagram representing a strong relationship, there is very little scatter about the line. This indicates, in the example shown on the chart, that hours studied is a good predictor of exam score.

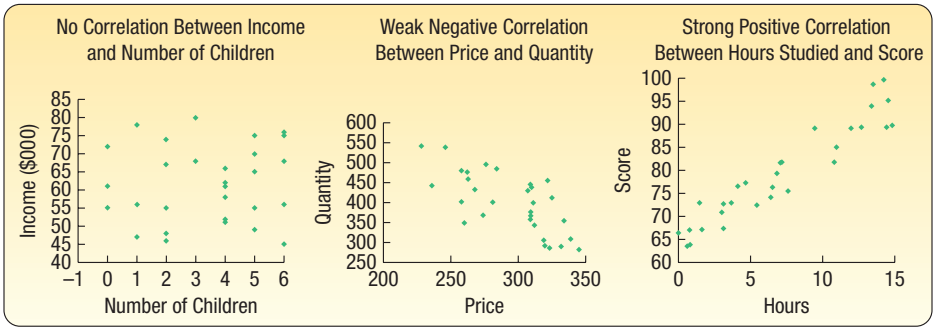
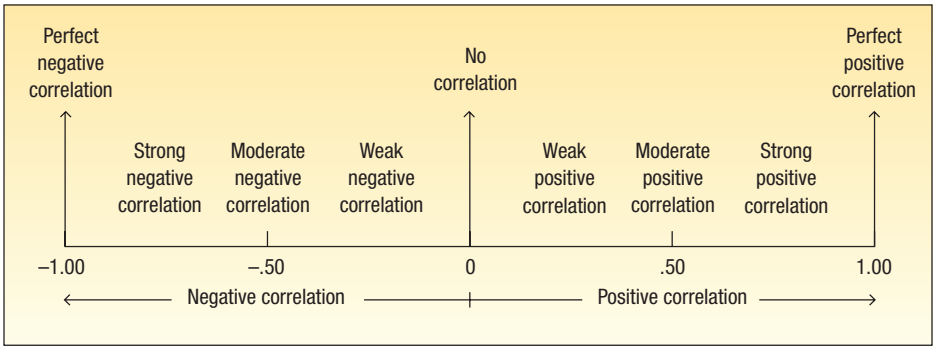


CHART 13–3 Scatter Diagrams Depicting Zero, Weak, and Strong Correlation

The following drawing summarizes the strength and direction of the correlation coefficient.



CORRELATION COEFFICIENT A measure of the strength of the linear relationship between two variables.

The characteristics of the correlation coefficient are summarized below.

CHARACTERISTICS OF THE CORRELATION COEFFICIENT

1. The sample correlation coefficient is identified by the lowercase letter *r*.
2. It shows the direction and strength of the linear relationship between two interval- or ratio-scale variables.
3. It ranges from −1 up to and including +1.
4. A value near 0 indicates there is little linear relationship between the variables.
5. A value near 1 indicates a direct or positive linear relationship between the variables.
6. A value near −1 indicates an inverse or negative linear relationship between the variables.

How is the value of the correlation coefficient determined? We will use the North American Copier Sales in Table 13–1 as an example. It is replicated in Table 13–2 for your convenience.

TABLE 13–2 Number of Sales Calls and Copiers Sold for 15 Salespeople

Sales Representative	Sales Calls	Copiers Sold
Brian Virost	96	41
Carlos Ramirez	40	41
Carol Saia	104	51
Greg Fish	128	60
Jeff Hall	164	61
Mark Reynolds	76	29
Meryl Rumsey	72	39
Mike Kiel	80	50
Ray Snarsky	36	28
Rich Niles	84	43
Ron Broderick	180	70
Sal Spina	132	56
Soni Jones	120	45
Susan Welch	44	31
Tom Keller	84	30
Total	1440	675

We begin with a scatter diagram, similar to Chart 13–2. Draw a vertical line through the data values at the mean of the *x*-values and a horizontal line at the mean of the *y*-values. In Chart 13–4, we've added a vertical line at 96 calls ($\bar{x} = \Sigma x/n = 1440/15 = 96$) and a horizontal line at 45 copiers ($\bar{y} = \Sigma y/n = 675/15 = 45$). These lines pass through the "center" of the data and divide the scatter diagram into four quadrants. Think of moving the origin from (0, 0) to (96, 45).

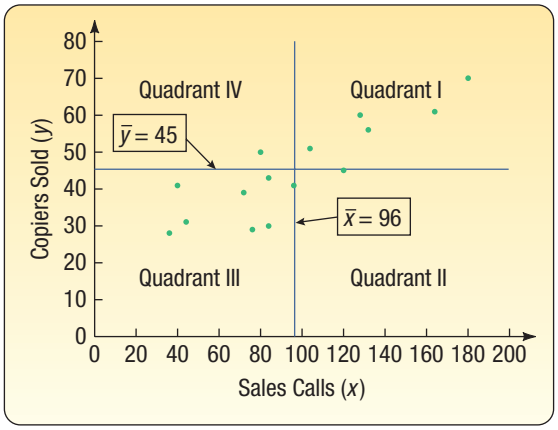


CHART 13–4 Computation of the Correlation Coefficient

Two variables are positively related when the number of copiers sold is above the mean and the number of sales calls is also above the mean. These points appear in the upper-right quadrant (labeled Quadrant I) of Chart 13–4. Similarly, when the number of copiers sold is less than the mean, so is the number of sales calls. These points fall in the lower-left quadrant of Chart 13–4 (labeled Quadrant III). For example, the third person on the list in Table 13–2, Carol Saia, made 104 sales calls and sold 51 copiers. These values are above their respective means, so this point is located in Quadrant I, which is in the upper-right quadrant. She made 8 more calls than the mean number of sales calls and sold 6 more than the mean number sold. Tom Keller, the last name on the list in Table 13–2, made 84 sales calls and sold 30 copiers. Both of these values are less than their respective means, hence this point is in the lower-left quadrant. Tom made 12 fewer sales calls and sold 15 fewer copiers than the respective means. The deviations from the mean number of sales calls and the mean number of copiers sold are summarized in Table 13–3 for the

TABLE 13–3 Deviations from the Mean and Their Products

Sales Representative	Sales Calls (x)	Copiers Sold (y)	$x - \bar{x}$	$x - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
Brian Virost	96	41	0	−4	0
Carlos Ramirez	40	41	−56	−4	224
Carol Saia	104	51	8	6	48
Greg Fish	128	60	32	15	480
Jeff Hall	164	61	68	16	1,088
Mark Reynolds	76	29	−20	−16	320
Meryl Rumsey	72	39	−24	−6	144
Mike Kiel	80	50	−16	5	−80
Ray Snarsky	36	28	−60	−17	1,020
Rich Niles	84	43	−12	−2	24
Ron Broderick	180	70	84	25	2,100
Sal Spina	132	56	36	11	396
Soni Jones	120	45	24	0	0
Susan Welch	44	31	−52	−14	728
Tom Keller	84	30	−12	−15	180
Totals	1440	675	0	0	6,672

15 sales representatives. The sum of the products of the deviations from the respective means is 6672. That is, the term $\Sigma(x - \bar{x})(y - \bar{y}) = 6672$.

In both the upper-right and the lower-left quadrants, the product of $(x - \bar{x})(y - \bar{y})$ is positive because both of the factors have the same sign. In our example, this happens for all sales representatives except Mike Kiel. Mike made 80 sales calls (which is less than the mean) but sold 50 machines (which is more than the mean). We can therefore expect the correlation coefficient to have a positive value.

If the two variables are inversely related, one variable will be above the mean and the other below the mean. Most of the points in this case occur in the upper-left and lower-right quadrants, that is, Quadrants II and IV. Now $(x - \bar{x})$ and $(y - \bar{y})$ will have opposite signs, so their product is negative. The resulting correlation coefficient is negative.

What happens if there is no linear relationship between the two variables? The points in the scatter diagram will appear in all four quadrants. The negative products of $(x - \bar{x})(y - \bar{y})$ offset the positive products, so the sum is near zero. This leads to a correlation coefficient near zero. So, the term $\Sigma(x - \bar{x})(y - \bar{y})$ drives the strength as well as the sign of the relationship between the two variables.

The correlation coefficient is also unaffected by the units of the two variables. For example, if we had used hundreds of copiers sold instead of the number sold, the correlation coefficient would be the same. The correlation coefficient is independent of the scale used if we divide the term $\Sigma(x - \bar{x})(y - \bar{y})$ by the sample standard deviations. It is also made independent of the sample size and bounded by the values +1.00 and −1.00 if we divide by $(n - 1)$.

This reasoning leads to the following formula:

CORRELATION COEFFICIENT

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} \qquad [13–1]$$

To compute the correlation coefficient, we use the standard deviations of the sample of 15 sales calls and 15 copiers sold. We could use formula (3–9) to calculate the sample standard deviations or we could use a statistical software package. For the specific Excel and Minitab commands, see the **Software Commands** in Appendix C. The following is the Excel output. The standard deviation of the number of sales calls is 42.76 and of the number of copiers sold 12.89.

	A	B	C	D	E	F	G	H
1		Sales Representative	Sales Calls (x)	Copiers Sold (y)			Sales Calls (x)	Copiers Sold (y)
2		Brian Virost	96	41		Mean	96.00	45.00
3		Carlos Ramirez	40	41		Standard Error	11.04	3.33
4		Carol Saia	104	51		Median	84.00	43.00
5		Greg Fish	128	60		Mode	84.00	41.00
6		Jeff Hall	164	61		Standard Deviation	42.76	12.89
7		Mark Reynolds	76	29		Sample Variance	1828.57	166.14
8		Meryl Rumsey	72	39		Kurtosis	-0.32	-0.73
9		Mike Kiel	80	50		Skewness	0.46	0.36
10		Ray Snarsky	36	28		Range	144.00	42.00
11		Rich Niles	84	43		Minimum	36.00	28.00
12		Ron Broderick	180	70		Maximum	180.00	70.00
13		Sal Spina	132	56		Sum	1440.00	675.00
14		Soni Jones	120	45		Count	15.00	15.00
15		Susan Welch	44	31				
16		Tom Keller	84	30				
17		Total	1440	675				

We now insert these values into formula (13–1) to determine the correlation coefficient:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} = \frac{6672}{(15 - 1)(42.76)(12.89)} = 0.865$$

How do we interpret a correlation of 0.865? First, it is positive, so we conclude there is a direct relationship between the number of sales calls and the number of copiers sold. This confirms our reasoning based on the scatter diagram, Chart 13–4. The value of 0.865 is fairly close to 1.00, so we conclude that the association is strong.

We must be careful with the interpretation. The correlation of 0.865 indicates a strong positive linear association between the variables. Ms. Bancer would be correct to encourage the sales personnel to make that extra sales call because the number of sales calls made is related to the number of copiers sold. However, does this mean that more sales calls *cause* more sales? No, we have not demonstrated cause and effect here, only that the two variables—sales calls and copiers sold—are statistically related.

If there is a strong relationship (say, .97) between two variables, we are tempted to assume that an increase or decrease in one variable *causes* a change in the other variable. For example, historically, the consumption of Georgia peanuts and the consumption of aspirin have a strong correlation. However, this does not indicate that an increase in the consumption of peanuts *caused* the consumption of aspirin to increase. Likewise, the incomes of professors and the number of inmates in mental institutions have increased proportionately. Further, as the population of donkeys has decreased, there has been an increase in the number of doctoral degrees granted. Relationships such as these are called **spurious correlations.** What we can conclude when we find two variables with a strong correlation is that there is a relationship or association between the two variables, not that a change in one causes a change in the other.

EXAMPLE

The Applewood Auto Group's marketing department believes younger buyers purchase vehicles on which lower profits are earned and the older buyers purchase vehicles on which higher profits are earned. They would like to use this information as part of an upcoming advertising campaign to try to attract older buyers, for whom the profits tend to be higher. Develop a scatter diagram depicting the relationship between vehicle profits and age of the buyer. Use statistical software to determine the correlation coefficient. Would this be a useful advertising feature?

SOLUTION

Using the Applewood Auto Group example, the first step is to graph the data using a scatter plot. It is shown in Chart 13–5.

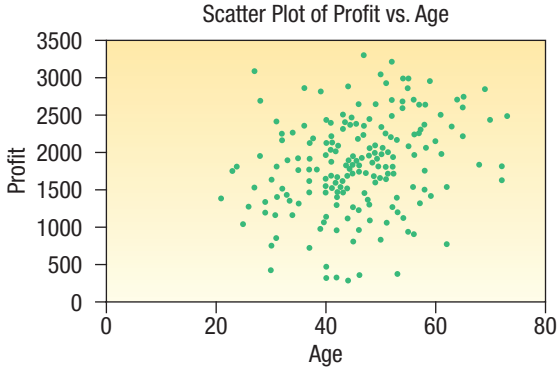


CHART 13–5 Scatter Diagram of Profit versus Age for the Applewood Auto Group Data

The scatter diagram suggests that a positive relationship does exist between age and profit; however, that relationship does not appear strong.

The next step is to calculate the correlation coefficient to evaluate the relative strength of the relationship. Statistical software provides an easy way to calculate the value of the correlation coefficient. The Excel output follows.

	A	B	C
1		***Age***	***Profit***
2	Age	1	
3	Profit	0.262	1

For these data, $r = 0.262$. To evaluate the relationship between a buyer's age and the profit on a car sale:

1. The relationship is positive or direct. Why? Because the sign of the correlation coefficient is positive. This confirms that as the age of the buyer increases, the profit on a car sale also increases.
2. The correlation coefficient is: $r = 0.262$. It is much closer to zero than one. Therefore, the relationship between the two variables is weak. We would observe that the relationship between the age of a buyer and the profit of their purchase is not very strong.

For Applewood Auto Group, the data does not support a business decision to create an advertising campaign to attract older buyers.

SELF-REVIEW 13–1

Haverty's Furniture is a family business that has been selling to retail customers in the Chicago area for many years. The company advertises extensively on radio, TV, and the Internet, emphasizing low prices and easy credit terms. The owner would like to review the relationship between sales and the amount spent on advertising. Below is information on sales and advertising expense for the last four months.

Month	Advertising Expense ($ million)	Sales Revenue ($ million)
July	2	7
August	1	3
September	3	8
October	4	10

(a) The owner wants to forecast sales on the basis of advertising expense. Which variable is the dependent variable? Which variable is the independent variable?
(b) Draw a scatter diagram.
(c) Determine the correlation coefficient.
(d) Interpret the strength of the correlation coefficient.

EXERCISES

1. FILE The following sample of observations were randomly selected.

x	4	5	3	6	10
y	4	6	5	7	7

Determine the correlation coefficient and interpret the relationship between x and y.

2. FILE The following sample of observations were randomly selected.

x	5	3	6	3	4	4	6	8
y	13	15	7	12	13	11	9	5

Determine the correlation coefficient and interpret the relationship between *x* and *y*.

3. FILE Bi-lo Appliance Super-Store has outlets in several large metropolitan areas in New England. The general sales manager aired a commercial for a digital camera on selected local TV stations prior to a sale starting on Saturday and ending Sunday. She obtained the information for Saturday–Sunday digital camera sales at the various outlets and paired it with the number of times the advertisement was shown on the local TV stations. The purpose is to find whether there is any relationship between the number of times the advertisement was aired and digital camera sales. The pairings are:

Location of TV Station	Number of Airings	Saturday–Sunday Sales ($ thousands)
Providence	4	15
Springfield	2	8
New Haven	5	21
Boston	6	24
Hartford	3	17

a. What is the dependent variable?
b. Draw a scatter diagram.
c. Determine the correlation coefficient.
d. Interpret these statistical measures.

4. FILE The production department of Celltronics International wants to explore the relationship between the number of employees who assemble a subassembly and the number produced. As an experiment, two employees were assigned to assemble the subassemblies. They produced 15 during a one-hour period. Then four employees assembled them. They produced 25 during a one-hour period. The complete set of paired observations follows.

Number of Assemblers	One-Hour Production (units)
2	15
4	25
1	10
5	40
3	30

The dependent variable is production; that is, it is assumed that different levels of production result from a different number of employees.
a. Draw a scatter diagram.
b. Based on the scatter diagram, does there appear to be any relationship between the number of assemblers and production? Explain.
c. Compute the correlation coefficient.

5. FILE The city council of Pine Bluffs is considering increasing the number of police in an effort to reduce crime. Before making a final decision, the council asked the chief of police to survey other cities of similar size to determine the relationship between the number of police and the number of crimes reported. The chief gathered the following sample information.

City	Police	Number of Crimes	City	Police	Number of Crimes
Oxford	15	17	Holgate	17	7
Starksville	17	13	Carey	12	21
Danville	25	5	Whistler	11	19
Athens	27	7	Woodville	22	6

a. Which variable is the dependent variable and which is the independent variable? Hint: Which of the following makes better sense: Cities with more police have fewer crimes, or cities with fewer crimes have more police? Explain your choice.
b. Draw a scatter diagram.
c. Determine the correlation coefficient.
d. Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?

6. FILE The owner of Maumee Ford-Volvo wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.

Car	Age (years)	Selling Price ($000)	Car	Age (years)	Selling Price ($000)
1	9	8.1	7	8	7.6
2	7	6.0	8	11	8.0
3	11	3.6	9	10	8.0
4	12	4.0	10	12	6.0
5	8	5.0	11	6	8.6
6	7	10.0	12	6	8.0

a. Draw a scatter diagram.
b. Determine the correlation coefficient.
c. Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?

Testing the Significance of the Correlation Coefficient

Recall that the sales manager of North American Copier Sales found the correlation between the number of sales calls and the number of copiers sold was 0.865. This indicated a strong positive association between the two variables. However, only 15 salespeople were sampled. Could it be that the correlation in the population is actually 0? This would mean the correlation of 0.865 was due to chance, or sampling error. The population in this example is all the salespeople employed by the firm.

Resolving this dilemma requires a test to answer the question: Could there be zero correlation in the population from which the sample was selected? To put it another way, did the computed r come from a population of paired observations with zero correlation? To continue our convention of allowing Greek letters to represent a population parameter, we will let ρ represent the correlation in the population. It is pronounced "rho."

We will continue with the illustration involving sales calls and copiers sold. We employ the same hypothesis testing steps described in Chapter 10. The null hypothesis and the alternate hypothesis are:

H_0: $\rho = 0$ (The correlation in the population is zero.)
H_1: $\rho \neq 0$ (The correlation in the population is different from zero.)

This is a two-tailed test. The null hypothesis can be rejected with either large or small sample values of the correlation coefficient.

The formula for t is:

t TEST FOR THE CORRELATION COEFFICIENT

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \text{ with } n-2 \text{ degrees of freedom} \qquad [13\text{–}2]$$

Using the .05 level of significance, the decision rule states that if the computed t falls in the area between plus 2.160 and minus 2.160, the null hypothesis is not rejected. To locate the critical value of 2.160, refer to Appendix B.5 for $df = n - 2 = 15 - 2 = 13$. See Chart 13–6.

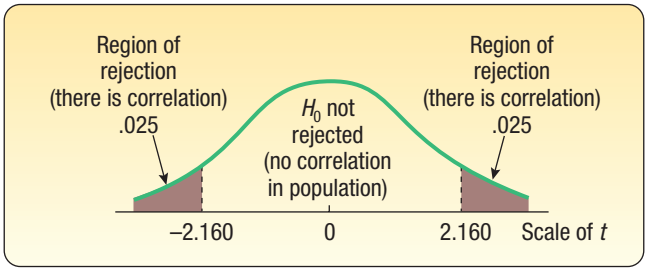


CHART 13–6 Decision Rule for Test of Hypothesis at .05 Significance Level and 13 *df*

Applying formula (13–2) to the example regarding the number of sales calls and units sold:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.865\sqrt{15-2}}{\sqrt{1-.865^2}} = 6.216$$

The computed t is in the rejection region. Thus, H_0 is rejected at the .05 significance level. Hence we conclude the correlation in the population is not zero. This indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of copiers sold in the population of salespeople.

We can also interpret the test of hypothesis in terms of p-values. A p-value is the likelihood of finding a value of the test statistic more extreme than the one computed, when H_0 is true. To determine the p-value, go to the t distribution in Appendix B.5 and find the row for 13 degrees of freedom. The value of the test statistic is 6.216, so in the row for 13 degrees of freedom and a two-tailed test, find the value closest to 6.216. For a two-tailed test at the 0.001 significance level, the critical value is 4.221. Because 6.216 is greater than 4.221, we conclude that the p-value is less than 0.001.

Both Minitab and Excel will report the correlation between two variables. In addition to the correlation, Minitab reports the p-value for the test of hypothesis that the correlation in the population between the two variables is 0. The Minitab output follows.

Worksheet 1 ***

	C1-T	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
	Name	Sales Calls	Copiers Sold								
1	Brian Virost	96	41								
2	Carlos Ramirez	40	41								
3	Carol Saia	104	51								
4	Greg Fish	128	60								
5	Jeff Hall	164	61								
6	Mark Reynolds	76	29								
7	Meryl Rumsey	72	39								
8	Mike Kiel	80	50								
9	Ray Snarsky	36	28								
10	Rich Niles	84	43								
11	Ron Broderick	180	70								
12	Sal Spina	132	56								
13	Soni Jones	120	45								
14	Susan Welch	44	31								
15	Tom Keller	84	30								

Session

```
Correlations: Sales Calls, Copiers Sold

Pearson correlation of Sales Calls and Copiers Sold = 0.865
P-Value = 0.000
```

EXAMPLE

In the Applewood Auto Group example on page 444, we found that the correlation coefficient between the profit on the sale of a vehicle by the Applewood Auto Group and the age of the person that purchased the vehicle was 0.262. The sign of the correlation coefficient was positive, so we concluded there was a direct relationship between the two variables. However, because the value of the correlation coefficient was small—that is, near zero—we concluded that an advertising campaign directed toward the older buyers was not warranted. We can test our conclusion by conducting a hypothesis test that the correlation coefficient is greater than zero using the .05 significance level.

SOLUTION

To test the hypothesis, we need to clarify the sample and population issues. Let's assume that the data collected on the 180 vehicles sold by the Applewood Group is a sample from the population of *all* vehicles sold over many years by the Applewood Auto Group. The Greek letter ρ is the correlation coefficient in the population and r the correlation coefficient in the sample.

Our next step is to set up the null hypothesis and the alternate hypothesis. We test the null hypothesis that the correlation coefficient is equal to or less than zero. The alternate hypothesis is that there is positive correlation between the two variables.

H_0: $\rho \leq 0$ (The correlation in the population is negative or equal to zero.)
H_1: $\rho > 0$ (The correlation in the population is positive.)

This is a one-tailed test because we are interested in confirming a positive association between the variables. The test statistic follows the t distribution with $n - 2$ degrees of freedom, so the degrees of freedom are $180 - 2 = 178$. However, the value for 178 degrees of freedom is not in Appendix B.5. The closest value is 180, so we will use that value. Our decision rule is to reject the null hypothesis if the computed value of the test statistic is greater than 1.653.

We use formula (13–2) to find the value of the test statistic.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.262\sqrt{180-2}}{\sqrt{1-0.262^2}} = 3.622$$

Comparing the value of our test statistic of 3.622 to the critical value of 1.653, we reject the null hypothesis. We conclude that the sample correlation coefficient of 0.262 is too large to have come from a population with no correlation. To put our results another way, there is a positive correlation between profits and age in the population.

This result is confusing and seems contradictory. On one hand, we observed that the correlation coefficient did not indicate a very strong relationship and that the Applewood Auto Group marketing department should not use this information for its promotion and advertising decisions. On the other hand, the hypothesis test indicated that the correlation coefficient is not equal to zero and that a positive relationship between age and profit exists. How can this be? We must be very careful about the application of the hypothesis test results. The hypothesis test shows a statistically significant result. However, this result does not necessarily support a practical decision to start a new marketing and promotion campaign to older purchasers. In fact, the relatively low correlation coefficient is an indication that the outcome of a new marketing and promotion campaign to older potential purchasers is, at best, uncertain.

SELF-REVIEW 13–2

A sample of 25 mayoral campaigns in medium-sized cities with populations between 50,000 and 250,000 showed that the correlation between the percent of the vote received and the amount spent on the campaign by the candidate was .43. At the .05 significance level, is there a positive association between the variables?

EXERCISES

7. The following hypotheses are given.

$$H_0: \rho \leq 0$$
$$H_1: \rho > 0$$

A random sample of 12 paired observations indicated a correlation of .32. Can we conclude that the correlation in the population is greater than zero? Use the .05 significance level.

8. The following hypotheses are given.

$$H_0: \rho \geq 0$$
$$H_1: \rho < 0$$

A random sample of 15 paired observations has a correlation of –.46. Can we conclude that the correlation in the population is less than zero? Use the .05 significance level.

9. Pennsylvania Refining Company is studying the relationship between the pump price of gasoline and the number of gallons sold. For a sample of 20 stations last Tuesday, the correlation was .78. At the .01 significance level, is the correlation in the population greater than zero?

10. A study of 20 worldwide financial institutions showed the correlation between their assets and pretax profit to be .86. At the .05 significance level, can we conclude that there is positive correlation in the population?

11. The Airline Passenger Association studied the relationship between the number of passengers on a particular flight and the cost of the flight. It seems logical that more passengers on the flight will result in more weight and more luggage, which in turn will result in higher fuel costs. For a sample of 15 flights, the correlation between the number of passengers and total fuel cost was .667. Is it reasonable to conclude that there is positive association in the population between the two variables? Use the .01 significance level.

12. FILE The Student Government Association at Middle Carolina University wanted to demonstrate the relationship between the number of beers a student drinks and his or her blood alcohol content (BAC). A random sample of 18 students participated in a study in which each participating student was randomly assigned a number of 12-ounce cans of beer to drink. Thirty minutes after they consumed their assigned number of beers, a member of the local sheriff's office measured their blood alcohol content. The sample information is reported below.

Student	Beers	BAC	Student	Beers	BAC
Charles	6	0.10	Jaime	3	0.07
Ellis	7	0.09	Shannon	3	0.05
Harriet	7	0.09	Nellie	7	0.08
Marlene	4	0.10	Jeanne	1	0.04
Tara	5	0.10	Michele	4	0.07
Kerry	3	0.07	Seth	2	0.06
Vera	3	0.10	Gilberto	7	0.12
Pat	6	0.12	Lillian	2	0.05
Marjorie	6	0.09	Becky	1	0.02

Use a statistical software package to answer the following questions.

a. Develop a scatter diagram for the number of beers consumed and BAC. Comment on the relationship. Does it appear to be strong or weak? Does it appear to be positive or inverse?

b. Determine the correlation coefficient.

c. At the .01 significance level, is it reasonable to conclude that there is a positive relationship in the population between the number of beers consumed and the BAC? What is the *p*-value?

LO13-3

Apply regression analysis to estimate the linear relationship between two variables.

STATISTICS IN ACTION

In finance, investors are interested in the trade-off between returns and risk. One technique to quantify risk is a regression analysis of a company's stock price (dependent variable) and an average measure of the stock market (independent variable). Often the Standard and Poor's (S&P) 500 Index is used to estimate the market. The regression coefficient, called beta in finance, shows the change in a company's stock price for a one-unit change in the S&P Index. For example, if a stock has a beta of 1.5, then when the S&P index increases by 1%, the stock price will increase by 1.5%. The opposite is also true. If the S&P decreases by 1%, the stock price will decrease by 1.5%. If the beta is 1.0, then a 1% change in the index should show a 1% change in a stock price. If the beta is less than 1.0, then a 1% change in the index shows less than a 1% change in the stock price.

REGRESSION ANALYSIS

In the previous sections of this chapter, we evaluated the direction and the significance of the linear relationship between two variables by finding the correlation coefficient. Regression analysis is another method to examine a linear relationship between two variables. This analysis uses the basic concepts of correlation but provides much more information by expressing the linear relationship between two variables in the form of an equation. Using this equation, we will be able to estimate the value of the dependent variable Y based on a selected value of the independent variable X. The technique used to develop the equation and provide the estimates is called **regression analysis.**

© Image Source/Getty Images RF

In Table 13–1, we reported the number of sales calls and the number of units sold for a sample of 15 sales representatives employed by North American Copier Sales. Chart 13–1 portrayed this information in a scatter diagram. Recall that we tested the significance of the correlation coefficient ($r = 0.865$) and concluded that a significant relationship exists between the two variables. Now we want to develop a linear equation that expresses the relationship between the number of sales calls, the independent variable, and the number of units sold, the dependent variable. The equation for the line used to estimate Y on the basis of X is referred to as the **regression equation.**

REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

Least Squares Principle

In regression analysis, our objective is to use the data to position a line that best represents the relationship between the two variables. Our first approach is to use a scatter diagram to visually position the line.

The scatter diagram in Chart 13–1 is reproduced in Chart 13–7, with a line drawn with a ruler through the dots to illustrate that a line would probably fit the data. However, the line drawn using a straight edge has one disadvantage: Its position is based in part on the judgment of the person drawing the line. The hand-drawn lines in Chart 13–8 represent the judgments of four people. All the lines except line *A* seem to be reasonable. That is, each line is centered among the graphed data. However, each would result in a different estimate of units sold for a particular number of sales calls.

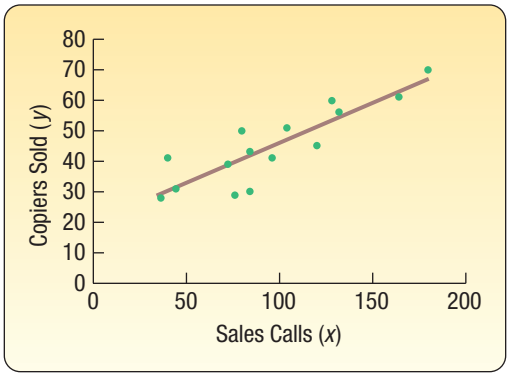


CHART 13–7 Sales Calls and Copiers Sold for 15 Sales Representatives

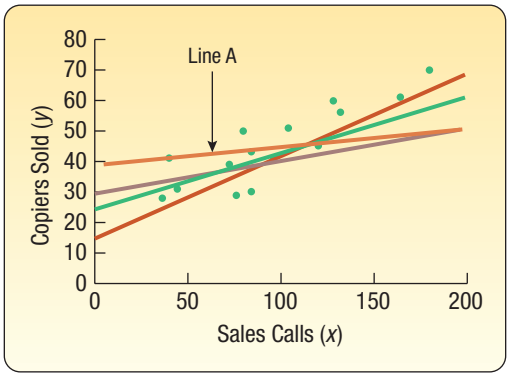


CHART 13–8 Four Lines Superimposed on the Scatter Diagram

We would prefer a method that results in a single, best regression line. This method is called the least squares principle. It gives what is commonly referred to as the "best-fitting" line.

LEAST SQUARES PRINCIPLE

A mathematical procedure that uses the data to position a line with the objective of minimizing the sum of the squares of the vertical distances between the actual y values and the predicted values of y.

To illustrate this concept, the same data are plotted in the three charts that follow. The dots are the actual values of y, and the asterisks are the predicted values of y for a given value of x. The regression line in Chart 13–9 was determined using the least squares method. It is the best-fitting line because the sum of the squares of the vertical deviations about it is at a minimum. The first plot ($x = 3$, $y = 8$) deviates by 2 from the line, found by $10 - 8$. The deviation squared is 4. The squared deviation for the plot $x = 4$, $y = 18$ is 16. The squared deviation for the plot $x = 5$, $y = 16$ is 4. The sum of the squared deviations is 24, found by $4 + 16 + 4$.

Assume that the lines in Charts 13–10 and 13–11 were drawn with a straight edge. The sum of the squared vertical deviations in Chart 13–10 is 44. For Chart 13–11, it is 132. Both sums are greater than the sum for the line in Chart 13–9, found by using the least squares method.

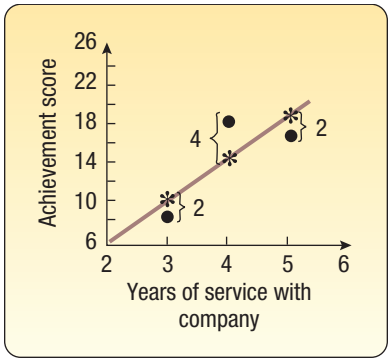


CHART 13–9 The Least Squares Line

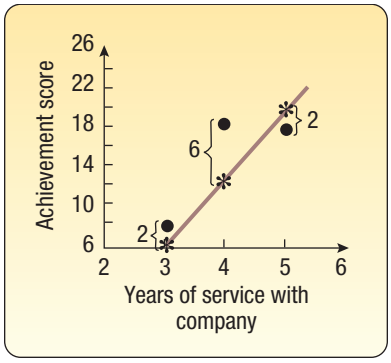


CHART 13–10 Line Drawn with a Straight Edge

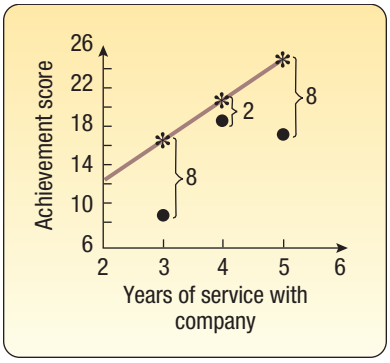


CHART 13–11 Different Line Drawn with a Straight Edge

The equation of a line has the form

GENERAL FORM OF LINEAR REGRESSION EQUATION

$$\hat{y} = a + bx \qquad [13–3]$$

where:

$\hat{y}$, read *y* hat, is the estimated value of the *y* variable for a selected *x* value.
a is the *y*-intercept. It is the estimated value of *Y* when $x = 0$. Another way to put it is: *a* is the estimated value of *y* where the regression line crosses the *Y*-axis when *x* is zero.
b is the slope of the line, or the average change in $\hat{y}$ for each change of one unit (either increase or decrease) in the independent variable *x*.
x is any value of the independent variable that is selected.

The general form of the linear regression equation is exactly the same form as the equation of any line. *a* is the *Y* intercept and *b* is the slope. The purpose of regression analysis is to calculate the values of *a* and *b* to develop a linear equation that best fits the data.

The formulas for *a* and *b* are:

SLOPE OF THE REGRESSION LINE

$$b = r\left(\frac{s_y}{s_x}\right) \qquad [13–4]$$

where:

r is the correlation coefficient.
s_y is the standard deviation of *y* (the dependent variable).
s_x is the standard deviation of *x* (the independent variable).

Y-INTERCEPT

$$a = \bar{y} - b\bar{x} \qquad [13–5]$$

where:

$\bar{y}$ is the mean of *y* (the dependent variable).
$\bar{x}$ is the mean of *x* (the independent variable).

EXAMPLE

Recall the example involving North American Copier Sales. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 15 sales representatives. As a part of her presentation at the upcoming sales meeting, Ms. Bancer, the sales manager, would like to offer specific information about the relationship between the number of sales calls and the number of copiers sold. Use the least squares method to determine a linear equation to express the relationship between the two variables. What is the expected number of copiers sold by a representative who made 100 calls?

SOLUTION

The first step in determining the regression equation is to find the slope of the least squares regression line. That is, we need the value of b. In the previous section on page 443, we determined the correlation coefficient r (0.865). In the Excel output on page 443, we determined the standard deviation of the independent variable x (42.76) and the standard deviation of the dependent variable y (12.89). The values are inserted in formula (13–4).

$$b = r\left(\frac{s_y}{s_x}\right) = .865\left(\frac{12.89}{42.76}\right) = 0.2608$$

Next, we need to find the value of a. To do this, we use the value for b that we just calculated as well as the means for the number of sales calls and the number of copiers sold. These means are also available in the Excel worksheet on page 443. From formula (13–5):

$$a = \bar{y} - b\bar{x} = 45 - .2608(96) = 19.9632$$

Thus, the regression equation is

$$\hat{y} = 19.9632 + 0.2608x.$$

So if a salesperson makes 100 calls, he or she can expect to sell 46.0432 copiers, found by

$$\hat{y} = 19.9632 + 0.2608x = 19.9632 + 0.2608(100) = 46.0432$$

The b value of .2608 indicates that for each additional sales call, the sales representative can expect to increase the number of copiers sold by about 0.2608. To put it another way, 20 additional sales calls in a month will result in about five more copiers being sold, found by 0.2608 (20) = 5.216.

The a value of 19.9632 is the point where the equation crosses the Y-axis. A literal translation is that if no sales calls are made, that is $x = 0$, 19.9632 copiers will be sold. Note that $x = 0$ is outside the range of values included in the sample and, therefore, should not be used to estimate the number of copiers sold. The sales calls ranged from 36 to 180, so estimates should be limited to that range.

Drawing the Regression Line

The least squares equation $\hat{y} = 19.9632 + 0.2608x$ can be drawn on the scatter diagram. The fifth sales representative in the sample is Jeff Hall. He made 164 calls. His estimated number of copiers sold is $\hat{y} = 19.9632 + 0.2608(164) = 62.7344$. The plot $x = 164$ and $\hat{y} = 62.7344$ is located by moving to 164 on the X-axis and then going vertically to 62.7344. The other points on the regression equation can be determined

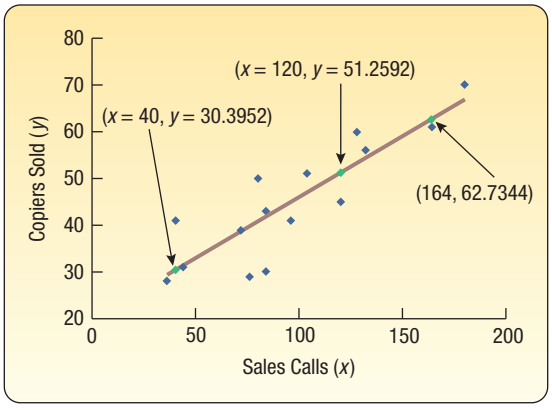


CHART 13–12 The Line of Regression Drawn on the Scatter Diagram

by substituting a particular value of x into the regression equation and calculating $\hat{y}$. All the points are connected to give the line. See Chart 13–12.

Sales Representative	Sales Calls (x)	Copiers Sold (y)	Estimated Sales ($\hat{y}$)
Brian Virost	96	41	45.0000
Carlos Ramirez	40	41	30.3952
Carol Saia	104	51	47.0864
Greg Fish	128	60	53.3456
Jeff Hall	164	61	62.7344
Mark Reynolds	76	29	39.7840
Meryl Rumsey	72	39	38.7408
Mike Kiel	80	50	40.8272
Ray Snarsky	36	28	29.3520
Rich Niles	84	43	41.8704
Ron Broderick	180	70	66.9072
Sal Spina	132	56	54.3888
Soni Jones	120	45	51.2592
Susan Welch	44	31	31.4384
Tom Keller	84	30	41.8704

The least squares regression line has some interesting and unique features. First, it will always pass through the point $(\bar{x}, \bar{y})$. To show this is true, we can use the mean number of sales calls to predict the number of copiers sold. In this example, the mean number of sales calls is 96, found by $\bar{x} = 1440/15$. The mean number of copiers sold is 45.0, found by $\bar{y} = 675/15$. If we let $x = 96$ and then use the regression equation to find the estimated value for $\hat{y}$ the result is:

$$\hat{y} = 19.9632 + 0.2608(96) = 45$$

The estimated number of copiers sold is exactly equal to the mean number of copiers sold. This simple example shows the regression line will pass through the point represented by the two means. In this case, the regression equation will pass through the point $x = 96$ and $y = 45$.

Second, as we discussed earlier in this section, there is no other line through the data where the sum of the squared deviations is smaller. To put it another way, the term $\Sigma(y - \hat{y})^2$ is smaller for the least squares regression equation than for any other equation. We use the Excel system to demonstrate this result in the following printout.

	A	B	C	D	E	F	G	H	I	J
1	**Sales Rep**	**Sales Calls (x)**	**Copiers Sold (y)**	**Estimated Sales**	$(y-\hat{y})$	$(y-\hat{y})^2$	y^*	$(y-y^*)^2$	y^{**}	$(y-y^{**})^2$
2	Brian Virost	96	41	45.0000	–4.0000	16.0000	44.4000	11.5600	41.6000	0.3600
3	Carlos Ramirez	40	41	30.3952	10.6048	112.4618	29.0000	144.0000	29.0000	144.0000
4	Carol Saia	104	51	47.0864	3.9136	15.3163	46.6000	19.3600	43.4000	57.7600
5	Greg Fish	128	60	53.3456	6.6544	44.2810	53.2000	46.2400	48.8000	125.4400
6	Jeff Hall	164	61	62.7344	–1.7344	3.0081	63.1000	4.4100	56.9000	16.8100
7	Mark Reynolds	76	29	39.7840	–10.7840	116.2947	38.9000	98.0100	37.1000	65.6100
8	Meryl Rumsey	72	39	38.7408	0.2592	0.0672	37.8000	1.4400	36.2000	7.8400
9	Mike Kiel	80	50	40.8272	9.1728	84.1403	40.0000	100.0000	38.0000	144.0000
10	Ray Snarsky	36	28	29.3520	–1.3520	1.8279	27.9000	0.0100	28.1000	0.0100
11	Rich Niles	84	43	41.8704	1.1296	1.2760	41.1000	3.6100	38.9000	16.8100
12	Ron Broderick	180	70	66.9072	3.0928	9.5654	67.5000	6.2500	60.5000	90.2500
13	Sal Spina	132	56	54.3888	1.6112	2.5960	54.3000	2.8900	49.7000	39.6900
14	Soni Jones	120	45	51.2592	–6.2592	39.1776	51.0000	36.0000	47.0000	4.0000
15	Susan Welch	44	31	31.4384	–0.4384	0.1922	30.1000	0.8100	29.9000	1.2100
16	Tom Keller	84	30	41.8704	–11.8704	140.9064	41.1000	123.2100	38.9000	79.2100
17	**Total**				**0.0000**	**587.1108**		**597.8000**		**793.0000**

In columns A, B, and C in the Excel spreadsheet above, we duplicated the sample information on sales and copiers sold from Table 13–1. In column D, we provide the estimated sales values, the $\hat{y}$ values, as calculated above.

In column E, we calculate the **residuals,** or the error values. This is the difference between the actual values and the predicted values. That is, column E is $(y - \hat{y})$. For Soni Jones,

$$\hat{y} = 19,9632 + 0.2608(120) = 51.2592$$

Her actual value is 45. So the residual, or error of estimate, is

$$(y - \hat{y}) = (45 - 51.2592) = -6.2592$$

This value reflects the amount the predicted value of sales is "off" from the actual sales value.

Next, in Column F, we square the residuals for each of the sales representatives and total the result. The total is 587.111.

$$\Sigma(y - \hat{y})^2 = 16.0000 + 112.4618 + \cdots + 140.9064 = 587.1108$$

This is the sum of the squared differences or the least squares value. There is no other line through these 15 data points where the sum of the squared differences is smaller.

We can demonstrate the least squares criterion by choosing two arbitrary equations that are close to the least squares equation and determining the sum of the squared differences for these equations. In column G, we use the equation $y^* = 18 + 0.275x$ to find the predicted value. Notice this equation is very similar to the least squares equation. In column H, we determine the residuals and square these residuals. For the first sales representative, Brian Virost,

$$y^* = 18 + 0.275(96) = 44.4$$

$$(y - y^*)^2 = (41 - 44.4)^2 = 11.56$$

This procedure is continued for the other 14 sales representatives and the squared residuals totaled. The result is 597.8. This is a larger value (597.8 is more than 587.1108) than the residuals for the least squares line.

In columns I and J on the output, we repeat the above process for yet another equation $y^{**} = 20 + 0.225x$. Again, this equation is similar to the least squares equation. The details for Brian Virost are:

$$y^{**} = 20 + 0.225x = 20 + 0.225(96) = 41.6$$

$$(y - y^{**})^2 = (41 - 41.6)^2 = 0.36$$

This procedure is continued for the other 14 sales representatives and the residuals totaled. The result is 793, which is also larger than the least squares values.

What have we shown with the example? The sum of the squared residuals $[\Sigma(y - \hat{y})^2]$ for the least squares equation is smaller than for other selected lines. The bottom line is you will not be able to find a line passing through these data points where the sum of the squared residuals is smaller.

SELF-REVIEW 13–3

Refer to Self-Review 13–1, where the owner of Haverty's Furniture Company was studying the relationship between sales and the amount spent on advertising. The advertising expense and sales revenue, both in millions of dollars, for the last 4 months are repeated below.

Month	Advertising Expense ($ million)	Sales Revenue ($ million)
July	2	7
August	1	3
September	3	8
October	4	10

(a) Determine the regression equation.
(b) Interpret the values of a and b.
(c) Estimate sales when $3 million is spent on advertising.

EXERCISES

13. FILE The following sample of observations was randomly selected.

x:	4	5	3	6	10
y:	4	6	5	7	7

a. Determine the regression equation.
b. Determine the value of when x is 7.

14. FILE The following sample of observations was randomly selected.

x:	5	3	6	3	4	4	6	8
y:	13	15	7	12	13	11	9	5

a. Determine the regression equation.
b. Determine the value of $\hat{y}$ when x is 7.

15. FILE Bradford Electric Illuminating Company is studying the relationship between kilowatt-hours (thousands) used and the number of rooms in a private single-family residence. A random sample of 10 homes yielded the following.

Number of Rooms	Kilowatt-Hours (thousands)	Number of Rooms	Kilowatt-Hours (thousands)
12	9	8	6
9	7	10	8
14	10	10	10
6	5	5	4
10	8	7	7

a. Determine the regression equation.
b. Determine the number of kilowatt-hours, in thousands, for a six-room house.

16. FILE Mr. James McWhinney, president of Daniel-James Financial Services, believes there is a relationship between the number of client contacts and the dollar amount of sales. To document this assertion, Mr. McWhinney gathered the following sample information. The *x* column indicates the number of client contacts last month and the *y* column shows the value of sales ($ thousands) last month for each client sampled.

Number of Contacts, *x*	Sales ($ thousands), *y*	Number of Contacts, *x*	Sales ($ thousands), *y*
14	24	23	30
12	14	48	90
20	28	50	85
16	30	55	120
46	80	50	110

a. Determine the regression equation.
b. Determine the estimated sales if 40 contacts are made.

17. FILE A recent article in *Bloomberg Businessweek* listed the "Best Small Companies." We are interested in the current results of the companies' sales and earnings. A random sample of 12 companies was selected and the sales and earnings, in millions of dollars, are reported below.

Company	Sales ($ millions)	Earnings ($ millions)	Company	Sales ($ millions)	Earnings ($ millions)
Papa John's International	$89.2	$4.9	Checkmate Electronics	$17.5	$ 2.6
Applied Innovation	18.6	4.4	Royal Grip	11.9	1.7
Integracare	18.2	1.3	M-Wave	19.6	3.5
Wall Data	71.7	8.0	Serving-N-Slide	51.2	8.2
Davidson & Associates	58.6	6.6	Daig	28.6	6.0
Chico's FAS	46.8	4.1	Cobra Golf	69.2	12.8

Let sales be the independent variable and earnings be the dependent variable.
a. Draw a scatter diagram.
b. Compute the correlation coefficient.
c. Determine the regression equation.
d. For a small company with $50.0 million in sales, estimate the earnings.

18. FILE We are studying mutual bond funds for the purpose of investing in several funds. For this particular study, we want to focus on the assets of a fund and its five-year performance. The question is: Can the five-year rate of return be estimated based on the assets of the fund? Nine mutual funds were selected at random, and their assets and rates of return are shown below.

Fund	Assets ($ millions)	Return (%)	Fund	Assets ($ millions)	Return (%)
AARP High Quality Bond	$622.2	10.8	MFS Bond A	$494.5	11.6
Babson Bond L	160.4	11.3	Nichols Income	158.3	9.5
Compass Capital Fixed Income	275.7	11.4	T. Rowe Price Short-term	681.0	8.2
Galaxy Bond Retail	433.2	9.1	Thompson Income B	241.3	6.8
Keystone Custodian B-1	437.9	9.2			

a. Draw a scatter diagram.
b. Compute the correlation coefficient.
c. Write a brief report of your findings for parts (a) and (b).
d. Determine the regression equation. Use assets as the independent variable.
e. For a fund with $400.0 million in sales, determine the five-year rate of return (in percent).

19. FILE Refer to Exercise 5. Assume the dependent variable is number of crimes.
a. Determine the regression equation.
b. Estimate the number of crimes for a city with 20 police officers.
c. Interpret the regression equation.

20. FILE Refer to Exercise 6.
a. Determine the regression equation.
b. Estimate the selling price of a 10-year-old car.
c. Interpret the regression equation.

LO13-4
Evaluate the significance of the slope of the regression equation.

TESTING THE SIGNIFICANCE OF THE SLOPE

In the prior section, we showed how to find the equation of the regression line that best fits the data. The method for finding the equation is based on the *least squares principle.* The purpose of the regression equation is to quantify a linear relationship between two variables.

The next step is to analyze the regression equation by conducting a test of hypothesis to see if the slope of the regression line is different from zero. Why is this important? If we can show that the slope of the line in the population is different from zero, then we can conclude that using the regression equation adds to our ability to predict or forecast the dependent variable based on the independent variable. If we cannot demonstrate that this slope is different from zero, then we conclude there is no merit to using the independent variable as a predictor. To put it another way, if we cannot show the slope of the line is different from zero, we might as well use the mean of the dependent variable as a predictor, rather than use the regression equation.

Following from the hypothesis-testing procedure in Chapter 10, the null and alternative hypotheses are:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

We use β (the Greek letter beta) to represent the population slope for the regression equation. This is consistent with our policy to identify population parameters by Greek letters. We assumed the information regarding North American Copier Sales, Table 13–2, is a sample. Be careful here. Remember, this is a single sample of salespeople, but when we selected a particular salesperson we identified two pieces of information: how many customers they called on and how many copiers they sold.

We identified the slope value as *b.* So *b* is our computed slope based on a sample and is an estimate of the population's slope, identified as β. The null hypothesis is that the slope of the regression equation in the population is zero. If this is the case, the regression line is horizontal and there is no relationship between the independent variable, *X,* and the dependent variable, *Y.* In other words, the value of the dependent variable is the same for any value of the independent variable and does not offer us any help in estimating the value of the dependent variable.

What if the null hypothesis is rejected? If the null hypothesis is rejected and the alternate hypothesis accepted, this indicates that the slope of the regression line for the population is not equal to zero. To put it another way, a significant relationship exists between the two variables. Knowing the value of the independent variable allows us to estimate the value of the dependent variable.

Before we test the hypothesis, we use statistical software to determine the needed regression statistics. We continue to use the North American Copier Sales data from Table 13–2 and use Excel to perform the necessary calculations. The following spreadsheet shows three tables to the right of the sample data.

	A	B	C	D	E	F	G	H	I	J
1	**Sales Representive**	**Sales calls (x)**	**Copiers Sold (y)**		SUMMARY OUTPUT					
2	Brian Virost	96	41							
3	Carlos Ramirez	40	41		*Regression Statistics*					
4	Carol Saia	104	51		Multiple R	0.865				
5	Greg Fish	128	60		R Square	0.748				
6	Jeff Hall	164	61		Adjusted R Square	0.728				
7	Mark Reynolds	76	29		Standard Error	6.720				
8	Meryl Rumsey	72	39		Observations	15				
9	Mike Kiel	80	50							
10	Ray Snarsky	36	28		ANOVA					
11	Rich Niles	84	43			*df*	*SS*	*MS*	*F*	*Significance F*
12	Ron Broderick	180	70		Regression	1	1738.89	1738.89	38.5031	3.19277E-05
13	Sal Spina	132	56		Residual	13	587.11	45.1623		
14	Sani Jones	120	45		Total	14	2326			
15	Susan Welch	44	31							
16	Tom Keller	84	30			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17					Intercept	19.9800	4.389675533	4.55159	0.00054	
18					Sales calls (x)	0.2606	0.042001817	6.20509	3.2E-05	

1. Starting on the top are the *Regression Statistics.* We will use this information later in the chapter, but notice that the "Multiple R" value is familiar. It is .865, which is the correlation coefficient we calculated using formula (13–1).
2. Next is an ANOVA table. This is a useful table for summarizing regression information. We will refer to it later in this chapter and use it extensively in the next chapter when we study multiple regression.
3. At the bottom, highlighted in blue, is the information needed to conduct our test of hypothesis regarding the slope of the line. It includes the value of the slope, which is 0.2606, and the intercept, which is 19.98. (Note that these values for the slope and the intercept are slightly different from those computed in the Example/Solution on page 454. These small differences are due to rounding.) In the column to the right of the regression coefficient is a column labeled "Standard Error." This is a value similar to the standard error of the mean. Recall that the standard error of the mean reports the variation in the sample means. In a similar fashion, these standard errors report the possible variation in slope and intercept values. The standard error of the slope coefficient is 0.0420.

To test the null hypothesis, we use the *t*-distribution with $(n - 2)$ and the following formula.

TEST FOR THE SLOPE $$t = \frac{b - 0}{s_b}$$ with $n - 2$ degrees of freedom **[13–6]**

where:

b is the estimate of the regression line's slope calculated from the sample information.

s_b is the standard error of the slope estimate, also determined from sample information.

Our first step is to set the null and the alternative hypotheses. They are:

$$H_0 : \beta \leq 0$$

$$H_1 : \beta > 0$$

Notice that we have a one-tailed test. If we do not reject the null hypothesis, we conclude that the slope of the regression line in the population could be zero. This means the independent variable is of no value in improving our estimate of the dependent

variable. In our case, this means that knowing the number of sales calls made by a representative does not help us predict the sales.

If we reject the null hypothesis and accept the alternative, we conclude the slope of the line is greater than zero. Hence, the independent variable is an aid in predicting the dependent variable. Thus, if we know the number of sales calls made by a salesperson, we can predict or forecast their sales. We also know, because we have demonstrated that the slope of the line is greater than zero—that is, positive—that more sales calls will result in the sale of more copiers.

The *t*-distribution is the test statistic; there are 13 degrees of freedom, found by $n - 2 = 15 - 2$. We use the .05 significance level. From Appendix B.5, the critical value is 1.771. Our decision rule is to reject the null hypothesis if the value computed from formula (13–6) is greater than 1.771. We apply formula (13–6) to find *t*.

$$t = \frac{b - 0}{s_b} = \frac{0.2606 - 0}{0.042} = 6.205$$

The computed value of 6.205 exceeds our critical value of 1.771, so we reject the null hypothesis and accept the alternative hypothesis. We conclude that the slope of the line is greater than zero. The independent variable, number of sales calls, is useful in estimating copier sales.

The table also provides us information on the *p*-value of this test. This cell is highlighted in purple. So we could select a significance level, say .05, and compare that value with the *p*-value. In this case, the calculated *p*-value in the table is reported in exponential notation and is equal to 0.0000319, so our decision is to reject the null hypothesis. An important caution is that the *p*-values reported in the statistical software are usually for a two-*tailed test.*

Before moving on, here is an interesting note. Observe that on page 448, when we conducted a test of hypothesis regarding the correlation coefficient for these same data using formula (13–2), we obtained the same value of the *t*-statistic, $t = 6.205$. Actually, when comparing the results of simple linear regression and correlation analysis, the two tests are equivalent and will always yield exactly the same values of *t* and the same *p*-values.

SELF-REVIEW 13–4

Refer to Self-Review 13–1, where the owner of Haverty's Furniture Company studied the relationship between the amount spent on advertising in a month and sales revenue for that month. The amount of sales is the dependent variable and advertising expense, the independent variable. The regression equation in that study was $\hat{y} = 1.5 + 2.2x$ for a sample of 5 months. Conduct a test of hypothesis to show there is a positive relationship between advertising and sales. From statistical software, the standard error of the regression coefficient is 0.42. Use the .05 significance level.

EXERCISES

21. FILE Refer to Exercise 5. The regression equation is $\hat{y} = 29.29 - 0.96x$, the sample size is 8, and the standard error of the slope is 0.22. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?

22. FILE Refer to Exercise 6. The regression equation is $\hat{y} = 11.18 - 0.49x$, the sample size is 12, and the standard error of the slope is 0.23. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?

23. FILE Refer to Exercise 17. The regression equation is $\hat{y} = 1.85 + .08x$, the sample size is 12, and the standard error of the slope is 0.03. Use the .05 significance level. Can we conclude that the slope of the regression line is *different from zero*?

24. FILE Refer to Exercise 18. The regression equation is $\hat{y} = 9.9198 - 0.00039x$, the sample size is 9, and the standard error of the slope is 0.0032. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?

LO13-5
Evaluate a regression equation's ability to predict using the standard estimate of the error and the coefficient of determination.

EVALUATING A REGRESSION EQUATION'S ABILITY TO PREDICT

The Standard Error of Estimate

The results of the regression analysis for North American Copier Sales show a significant relationship between number of sales calls and the number of sales made. By substituting the names of the variables into the equation, it can be written as:

$$\text{Number of copiers sold} = 19.9632 + 0.2608\ (\text{Number of sales calls})$$

The equation can be used to estimate the number of copiers sold for any given "number of sales calls" within the range of the data. For example, if the number of sales calls is 84, then we can predict the number of copiers sold. It is 41.8704, found by 19.9632 + 0.2608(84). However, the data show two sales representatives with 84 sales calls and 30 and 43 copiers sold. So, is the regression equation a good predictor of "Number of copiers sold"?

Perfect prediction, which is finding the exact outcome, is practically impossible in almost all disciplines including economics and business. For example:

- A large electronics firm, with production facilities throughout the United States, has a stock option plan for employees. Suppose there is a relationship between the number of years employed and the number of shares owned. This relationship is likely because, as number of years of service increases, the number of shares an employee earns also increases. If we observe all employees with 20 years of service, they would most likely own different numbers of shares.
- A real estate developer in the southwest United States studied the relationship between the income of buyers and the size, in square feet, of the home they purchased. The developer's analysis shows that as the income of a purchaser increases, the size of the home purchased will also increase. However, all buyers with an income of $70,000 will not purchase a home of exactly the same size.

What is needed, then, is a measure that describes how precise the prediction of Y is based on X or, conversely, how inaccurate the estimate might be. This measure is called the **standard error of estimate.** The standard error of estimate is symbolized by $s_{y \cdot x}$. The subscript, $y \cdot x$, is interpreted as the standard error of y for a given value of x. It is the same concept as the standard deviation discussed in Chapter 3. The standard deviation measures the dispersion around the mean. The standard error of estimate measures the dispersion about the regression line for a given value of x.

STANDARD ERROR OF ESTIMATE A measure of the dispersion, or scatter, of the observed values around the line of regression for a given value of x.

The standard error of estimate is found using formula (13–7).

STANDARD ERROR OF ESTIMATE

$$s_{y \cdot x} = \sqrt{\frac{\Sigma (y - \hat{y})^2}{n - 2}} \qquad [13\text{–}7]$$

The calculation of the standard error of estimate requires the sum of the squared differences between each observed value of y and the predicted value of y, which is identified as $\hat{y}$ in the numerator. This calculation is illustrated in the following spreadsheet. See the highlighted cell in the lower right corner.

	A	B	C	D	E	F
1	**Sales Rep**	**Sales Calls (x)**	**Copiers Sold (y)**	**Estimated Sales**	$(y-\hat{y})$	$(y-\hat{y})^2$
2	Brian Virost	96	41	45.0000	–4.0000	16.0000
3	Carlos Ramirez	40	41	30.3952	10.6048	112.4618
4	Carol Saia	104	51	47.0864	3.9136	15.3163
5	Greg Fish	128	60	53.3456	6.6544	44.2810
6	Jeff Hall	164	61	62.7344	–1.7344	3.0081
7	Mark Reynolds	76	29	39.7840	–10.7840	116.2947
8	Meryl Rumsey	72	39	38.7408	0.2592	0.0672
9	Mike Kiel	80	50	40.8272	9.1728	84.1403
10	Ray Snarsky	36	28	29.3520	–1.3520	1.8279
11	Rich Niles	84	43	41.8704	1.1296	1.2760
12	Ron Broderick	180	70	66.9072	3.0928	9.5654
13	Sal Spina	132	56	54.3888	1.6112	2.5960
14	Soni Jones	120	45	51.2592	–6.2592	39.1776
15	Susan Welch	44	31	31.4384	–0.4384	0.1922
16	Tom Keller	84	30	41.8704	–11.8704	140.9064
17	**Total**				**0.0000**	**587.1108**

The calculation of the standard error of estimate is:

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{587.1108}{15 - 2}} = 6.720$$

The standard error of estimate can be calculated using statistical software such as Excel. It is included in Excel's regression analysis on page 460 and highlighted in yellow. Its value is 6.720.

If the standard error of estimate is small, this indicates that the data are relatively close to the regression line and the regression equation can be used to predict *y* with little error. If the standard error of estimate is large, this indicates that the data are widely scattered around the regression line and the regression equation will not provide a precise estimate of *y*.

The Coefficient of Determination

Using the standard error of estimate provides a relative measure of a regression equation's ability to predict. We will use it to provide more specific information about a prediction in the next section. In this section, another statistic is explained that will provide a more interpretable measure of a regression equation's ability to predict. It is called the coefficient of determination, or *R*-square.

COEFFICIENT OF DETERMINATION

The proportion of the total variation in the dependent variable *Y* that is explained, or accounted for, by the variation in the independent variable *X*.

The coefficient of determination is easy to compute. It is the correlation coefficient squared. Therefore, the term *R*-square is also used. With the North American Copier Sales data, the correlation coefficient for the relationship between the number of copiers sold and the number of sales calls is 0.865. If we compute $(0.865)^2$, the coefficient of determination is 0.748. See the blue (Multiple R) and green (*R*-square) highlighted cells in the spreadsheet on page 460. To better interpret the coefficient of determination, convert it to a percentage. Hence, we say that 74.8% of the variation in the number of copiers sold is explained, or accounted for, by the variation in the number of sales calls.

How well can the regression equation predict number of copiers sold with number of sales calls made? If it were possible to make perfect predictions, the coefficient of determination would be 100%. That would mean that the independent variable, number of sales calls, explains or accounts for all the variation in the number of copiers sold. A coefficient of determination of 100% is associated with a correlation coefficient of +1.0 or −1.0. Refer to Chart 13–2, which shows that a perfect prediction is associated with a perfect linear relationship where all the data points form a perfect line in a scatter diagram. Our analysis shows that only 74.8% of the variation in copiers sold is explained by the number of sales calls. Clearly, these data do not form a perfect line. Instead, the data are scattered around the best-fitting, least squares regression line, and there will be error in the predictions. In the next section, the standard error of estimate is used to provide more specific information regarding the error associated with using the regression equation to make predictions.

SELF-REVIEW 13–5

Refer to Self-Review 13–1, where the owner of Haverty's Furniture Company studied the relationship between the amount spent on advertising in a month and sales revenue for that month. The amount of sales is the dependent variable and advertising expense is the independent variable.

(a) Determine the standard error of estimate.
(b) Determine the coefficient of determination.
(c) Interpret the coefficient of determination.

EXERCISES

(You may wish to use a statistical software package such as Excel, Minitab, or Megastat to assist in your calculations.)

25. Refer to Exercise 5. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
26. Refer to Exercise 6. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
27. Refer to Exercise 15. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
28. Refer to Exercise 16. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.

Relationships among the Correlation Coefficient, the Coefficient of Determination, and the Standard Error of Estimate

In formula 13-7 shown on page 462, we described the standard error of estimate. Recall that it measures how close the actual values are to the regression line. When the standard error is small, it indicates that the two variables are closely related. In the calculation of the standard error, the key term is

$$\Sigma(y - \hat{y})^2$$

If the value of this term is small, then the standard error will also be small.

The correlation coefficient measures the strength of the linear association between two variables. When the points on the scatter diagram appear close to the line, we note that the correlation coefficient tends to be large. Therefore, the correlation coefficient and the standard error of the estimate are inversely related. As the strength of a linear relationship between two variables increases, the correlation coefficient increases and the standard error of the estimate decreases.

We also noted that the square of the correlation coefficient is the coefficient of determination. The coefficient of determination measures the percentage of the variation in *Y* that is explained by the variation in *X*.

A convenient vehicle for showing the relationship among these three measures is an ANOVA table. See the highlighted portion of the spreadsheet below. This table is similar to the analysis of variance table developed in Chapter 12. In that chapter, the total variation was divided into two components: variation due to the *treatments* and variation due to *random error.* The concept is similar in regression analysis. The total variation is divided into two components: (1) variation explained by the *regression* (explained by the independent variable) and (2) the *error,* or *residual.* This is the unexplained variation. These three sources of variance (total, regression, and residual) are identified in the first column of the spreadsheet ANOVA table. The column headed "*df*" refers to the degrees of freedom associated with each category. The total number of degrees of freedom is $n - 1$. The number of degrees of freedom in the regression is 1 because there is only one independent variable. The number of degrees of freedom associated with the error term is $n - 2$. The term SS located in the middle of the ANOVA table refers to the sum of squares. You should note that the total degrees of freedom are equal to the sum of the regression and residual (error) degrees of freedom, and the total sum of squares is equal to the sum of the regression and residual (error) sum of squares. This is true for any ANOVA table.

	A	B	C	D	E	F	G	H	I	J
1	**Sales Representive**	**Sales Calls (x)**	**Copiers Sold (y)**		**SUMMARY OUTPUT**					
2	Brian Virost	96	41		***Regression Statistics***					
3	Carlos Ramirez	40	41		Multiple R	0.865				
4	Carol Saia	104	51		R Square	0.748				
5	Greg Fish	128	60		Adjusted R Square	0.728				
6	Jeff Hall	164	61		Standard Error	6.720				
7	Mark Reynolds	76	29		Observations	15				
8	Meryl Rumsey	72	39							
9	Mike Kiel	80	50		**ANOVA**					
10	Ray Snarsky	36	28			**df**	**SS**	**MS**	**F**	**Significance F**
11	Rich Niles	84	43		Regression	1	1738.890	1738.890	38.503	0.000
12	Ron Broderick	180	70		Residual	13	587.110	45.162		
13	Sal Spina	132	56		Total	14	2326			
14	Sani Jones	120	45							
15	Susan Welch	44	31			***Coefficients***	***Standard Error***	***t Stat***	***P-value***	
16	Tom Keller	84	30		Intercept	19.980	4.390	4.552	0.001	
17					Sales Calls (x)	0.261	0.042	6.205	0.000	

The ANOVA sum of squares are:

$$\text{Regression Sum of Squares} = \text{SSR} = \Sigma(\hat{y} - \bar{y})^2 = 1738.89$$

$$\text{Residual or Error Sum of Squares} = \text{SSE} = \Sigma(y - \hat{y})^2 = 587.11$$

$$\text{Total Sum of Squares} = \text{SS Total} = \Sigma(y - \bar{y})^2 = 2326.0$$

Recall that the coefficient of determination is defined as the percentage of the total variation (SS Total) explained by the regression equation (SSR). Using the ANOVA table, the reported value of *R*-square can be validated.

COEFFICIENT OF DETERMINATION

$$r^2 = \frac{\text{SSR}}{\text{SS Total}} = 1 - \frac{\text{SSE}}{\text{SS Total}} \qquad [13\text{–}8]$$

Using the values from the ANOVA table, the coefficient of determination is 1738.89/2326.0 = 0.748. Therefore, the more variation of the dependent variable (SS Total) explained by the independent variable (SSR), the higher the coefficient of determination.

We can also express the coefficient of determination in terms of the error or residual variation:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS Total}} = 1 - \frac{587.11}{2326.0} = 1 - 0.252 = 0.748$$

As illustrated in formula 13–8, the coefficient of determination and the residual or error sum of squares are inversely related. The higher the unexplained or error variation as a percentage of the total variation, the lower is the coefficient of determination. In this case, 25.2% of the total variation in the dependent variable is error or residual variation.

The final observation that relates the correlation coefficient, the coefficient of determination, and the standard error of estimate is to show the relationship between the standard error of estimate and SSE. By substituting [SSE Residual or Error Sum of Squares = SSE = $\Sigma(y - \hat{y})^2$] into the formula for the standard error of estimate, we find:

STANDARD ERROR OF ESTIMATE

$$s_{y \cdot x} = \sqrt{\frac{\text{SSE}}{n - 2}} \qquad \textbf{[13–9]}$$

Note that $s_{y \cdot x}$ can also be computed using the residual mean square from the ANOVA table.

STANDARD ERROR OF THE ESTIMATE

$$s_{y \cdot x} = \sqrt{\textit{Residual mean square}} \qquad \textbf{[13–10]}$$

In sum, regression analysis provides two statistics to evaluate the predictive ability of a regression equation: the standard error of the estimate and the coefficient of determination. When reporting the results of a regression analysis, the findings must be clearly explained, especially when using the results to make predictions of the dependent variable. The report must always include a statement regarding the coefficient of determination so that the relative precision of the prediction is known to the reader of the report. Objective reporting of statistical analysis is required so that the readers can make their own decisions.

EXERCISES

29. Given the following ANOVA table:

Source	*DF*	SS	MS	*F*
Regression	1	1000.0	1000.0	26.00
Error	13	500.0	38.46	
Total	14	1500.0		

a. Determine the coefficient of determination.
b. Assuming a direct relationship between the variables, what is the correlation coefficient?
c. Determine the standard error of estimate.

30. On the first statistics exam, the coefficient of determination between the hours studied and the grade earned was 80%. The standard error of estimate was 10. There were 20 students in the class. Develop an ANOVA table for the regression analysis of hours studied as a predictor of the grade earned on the first statistics exam.

LO13-6
Calculate and interpret confidence and prediction intervals.

STATISTICS IN ACTION

Studies indicate that for both men and women, those who are considered good looking earn higher wages than those who are not. In addition, for men there is a correlation between height and salary. For each additional inch of height, a man can expect to earn an additional $250 per year. So a man 6′6″ tall receives a $3,000 "stature" bonus over his 5′6″ counterpart. Being overweight or underweight is also related to earnings, particularly among women. A study of young women showed the heaviest 10% earned about 6% less than their lighter counterparts.

INTERVAL ESTIMATES OF PREDICTION

The standard error of estimate and the coefficient of determination are two statistics that provide an overall evaluation of the ability of a regression equation to predict a dependent variable. Another way to report the ability of a regression equation to predict is specific to a stated value of the independent variable. For example, we can predict the number of copiers sold (y) for a selected value of number of sales calls made (x). In fact, we can calculate a confidence interval for the predicted value of the dependent variable for a selected value of the independent variable.

Assumptions Underlying Linear Regression

Before we present the confidence intervals, the assumptions for properly applying linear regression should be reviewed. Chart 13–13 illustrates these assumptions.

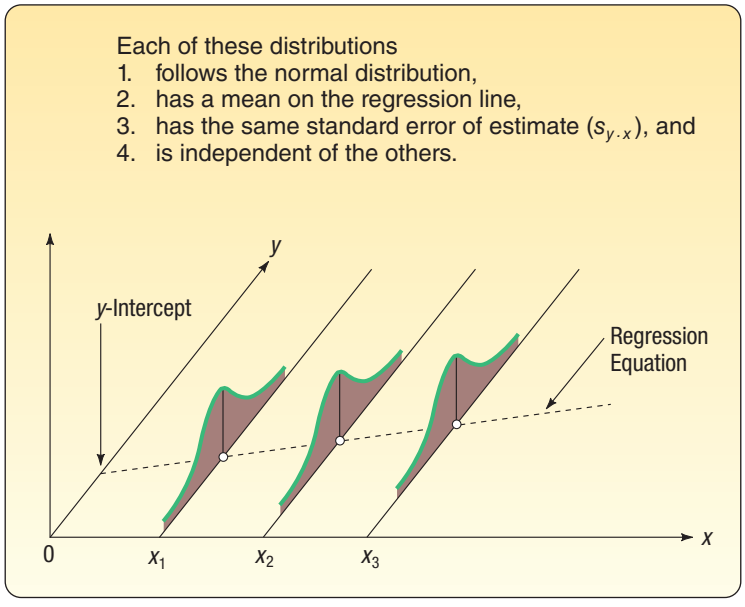


CHART 13–13 Regression Assumptions Shown Graphically

1. For each value of x, there are corresponding y values. These y values follow the normal distribution.
2. The means of these normal distributions lie on the regression line.
3. The standard deviations of these normal distributions are all the same. The best estimate we have of this common standard deviation is the standard error of estimate ($s_{y \cdot x}$).
4. The y values are statistically independent. This means that in selecting a sample, a particular x does not depend on any other value of x. This assumption is particularly important when data are collected over a period of time. In such situations, the errors for a particular time period are often correlated with those of other time periods.

Recall from Chapter 7 that if the values follow a normal distribution, then the mean plus or minus one standard deviation will encompass 68% of the observations, the mean plus or minus two standard deviations will encompass 95% of the observations, and

the mean plus or minus three standard deviations will encompass virtually all of the observations. The same relationship exists between the predicted values $\hat{y}$ and the standard error of estimate ($s_{y \cdot x}$).

1. $\hat{y} \pm s_{y \cdot x}$ will include the middle 68% of the observations.
2. $\hat{y} \pm 2s_{y \cdot x}$ will include the middle 95% of the observations.
3. $\hat{y} \pm 3s_{y \cdot x}$ will include virtually all the observations.

We can now relate these assumptions to North American Copier Sales, where we studied the relationship between the number of sales calls and the number of copiers sold. If we drew a parallel line 6.72 units above the regression line and another 6.72 units below the regression line, about 68% of the points would fall between the two lines. Similarly, a line 13.44 [$2s_{y \cdot x} = 2(6.72)$] units above the regression line and another 13.44 units below the regression line should include about 95% of the data values.

As a rough check, refer to column E in the Excel spreadsheet appearing on page 456. Four of the 15 deviations exceed one standard error of estimate. That is, the deviations of Carlos Ramirez, Mark Reynolds, Mike Keil, and Tom Keller all exceed 6.72 (one standard error). All values are less than 13.44 units away from the regression line. In short, 11 of the 15 deviations are within one standard error and all are within two standard errors. That is a fairly good result for a relatively small sample.

Constructing Confidence and Prediction Intervals

When using a regression equation, two different predictions can be made for a selected value of the independent variable. The differences are subtle but very important and are related to the assumptions stated in the last section. Recall that for any selected value of the independent variable (X), the dependent variable (Y) is a random variable that is normally distributed with a mean $\hat{Y}$. Each distribution of Y has a standard deviation equal to the regression analysis's standard error of estimate.

The first interval estimate is called a **confidence interval.** This is used when the regression equation is used to predict the mean value of Y for a given value of x. For example, we would use a confidence interval to estimate the mean salary of all executives in the retail industry based on their years of experience. To determine the confidence interval for the mean value of y for a given x, the formula is:

CONFIDENCE INTERVAL FOR THE MEAN OF Y, GIVEN X

$$\hat{y} \pm ts_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma (x - \bar{x})^2}} \qquad [13\text{–}11]$$

The second interval estimate is called a prediction interval. This is used when the regression equation is used to predict an individual y for a given value of x. For example, we would estimate the salary of a particular retail executive who has 20 years of experience. To calculate a prediction interval, formula 13-11 is modified by adding a 1 under the radical. To determine the prediction interval for an estimate of an individual for a given x, the formula is:

PREDICTION INTERVAL FOR Y, GIVEN X

$$\hat{y} \pm ts_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma (x - \bar{x})^2}} \qquad [13\text{–}12]$$

EXAMPLE

We return to the North American Copier Sales illustration. Determine a 95% confidence interval for all sales representatives who make 50 calls, and determine a prediction interval for Sheila Baker, a West Coast sales representative who made 50 calls.

SOLUTION

We use formula (13–11) to determine a confidence level. Table 13–4 includes the necessary totals and a repeat of the information of Table 13–2.

TABLE 13–4 Determining Confidence and Prediction Intervals

Sales Representative	Sales Calls (x)	Copiers Sold (y)	$(x - \bar{x})$	$(x - \bar{x})^2$
Brian Virost	96	41	0	0
Carlos Ramirez	40	41	−56	3,136
Carol Saia	104	51	8	64
Greg Fish	128	60	32	1,024
Jeff Hall	164	61	68	4,624
Mark Reynolds	76	29	−20	400
Meryl Rumsey	72	39	−24	576
Mike Kiel	80	50	−16	256
Ray Snarsky	36	28	−60	3,600
Rich Niles	84	43	−12	144
Ron Broderick	180	70	84	7,056
Sal Spina	132	56	36	1,296
Sani Jones	120	45	24	576
Susan Welch	44	31	−52	2,704
Tom Keller	84	30	−12	144
Total	1440	675	0	25,600

The first step is to determine the number of copiers we expect a sales representative to sell if he or she makes 50 calls. It is 33.0032, found by

$$\hat{y} = 19.9632 + 0.2608x = 19.9632 + 0.2608\,(50) = 33.0032$$

To find the t value, we need to first know the number of degrees of freedom. In this case, the degrees of freedom are $n - 2 = 15 - 2 = 13$. We set the confidence level at 95%. To find the value of t, move down the left-hand column of Appendix B.5 to 13 degrees of freedom, then move across to the column with the 95% level of confidence. The value of t is 2.160.

In the previous section, we calculated the standard error of estimate to be 6.720. We let $x = 50$, and from Table 13–4, the mean number of sales calls is 96.0 (1440/15) and $\Sigma(x - \bar{x})^2 = 25,600$ Inserting these values in formula (13–11), we can determine the confidence interval.

$$\text{Confidence Interval} = \hat{y} \pm ts_{y\cdot x}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}}$$

$$= 33.0032 \pm 2.160(6.720)\sqrt{\frac{1}{15} + \frac{(50 - 96)^2}{25,600}}$$

$$= 33.0032 \pm 5.6090$$

Thus, the 95% confidence interval for all sales representatives who make 50 calls is from 27.3942 up to 38.6122. To interpret, let's round the values. If a sales representative makes 50 calls, he or she can expect to sell 33 copiers. It is likely the sales will range from 27.4 to 38.6 copiers.

Suppose we want to estimate the number of copiers sold by Sheila Baker, who made 50 sales calls. Using formula (13–12), the 95% prediction interval is determined as follows:

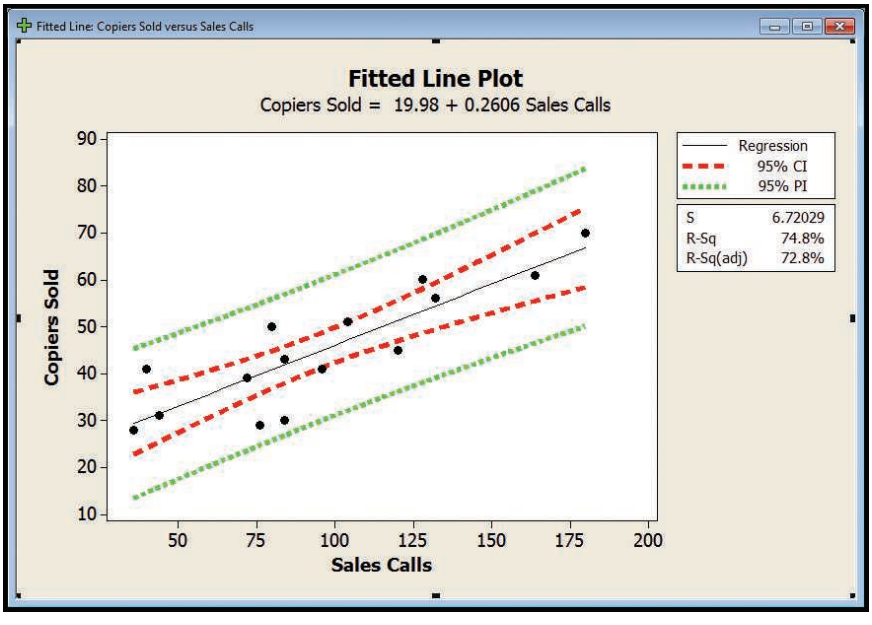
$$\text{Prediction Interval} = \hat{y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}}$$

$$= 33.0032 \pm 2.160(6.720)\sqrt{1 + \frac{1}{15} + \frac{(50 - 96)^2}{25{,}600}}$$

$$= 33.0032 \pm 15.5612$$

Thus, the interval is from 17.442 up to 48.5644 copiers. We conclude that the number of office machines sold will be between about 17.4 and 48.6 for a particular sales representative, such as Sheila Baker, who makes 50 calls. This interval is quite large. It is much larger than the confidence interval for all sales representatives who made 50 calls. It is logical, however, that there should be more variation in the sales estimate for an individual than for a group.

The following Minitab graph shows the relationship between the least squares regression line (in the center), the confidence interval (shown in crimson), and the prediction interval (shown in green). The bands for the prediction interval are always further from the regression line than those for the confidence interval. Also, as the values of x move away from the mean number of calls (96) in either direction, the confidence interval and prediction interval bands widen. This is caused by the numerator of the right-hand term under the radical in formulas (13–11) and (13–12). That is, as the term increases, the widths of the confidence interval and the prediction interval also increase. To put it another way, there is less precision in our estimates as we move away, in either direction, from the mean of the independent variable.



We wish to emphasize again the distinction between a confidence interval and a prediction interval. A confidence interval refers to the mean of all cases for a given value of x and is computed by formula (13–11). A prediction interval refers to a particular, single case for a given value of x and is computed using formula (13–12). The prediction interval will always be wider because of the extra 1 under the radical in the second equation.

SELF-REVIEW 13–6

Refer to the sample data in Self-Review 13–1, where the owner of Haverty's Furniture was studying the relationship between sales and the amount spent on advertising. The advertising expense and sales revenue, both in millions of dollars, for the last 4 months are repeated below.

Month	Advertising Expense ($ million)	Sales Revenue ($ million)
July	2	7
August	1	3
September	3	8
October	4	10

The regression equation was computed to be $\hat{y} = 1.5 + 2.2x$ and the standard error 0.9487. Both variables are reported in millions of dollars. Determine the 90% confidence interval for the typical month in which $3 million was spent on advertising.

EXERCISES

31. Refer to Exercise 13.
- **a.** Determine the .95 confidence interval for the mean predicted when $x = 7$.
- **b.** Determine the .95 prediction interval for an individual predicted when $x = 7$.

32. Refer to Exercise 14.
- **a.** Determine the .95 confidence interval for the mean predicted when $x = 7$.
- **b.** Determine the .95 prediction interval for an individual predicted when $x = 7$.

33. Refer to Exercise 15.
- **a.** Determine the .95 confidence interval, in thousands of kilowatt-hours, for the mean of all six-room homes.
- **b.** Determine the .95 prediction interval, in thousands of kilowatt-hours, for a particular six-room home.

34. Refer to Exercise 16.
- **a.** Determine the .95 confidence interval, in thousands of dollars, for the mean of all sales personnel who make 40 contacts.
- **b.** Determine the .95 prediction interval, in thousands of dollars, for a particular salesperson who makes 40 contacts.

TRANSFORMING DATA

LO13-7
Use a log function to transform a nonlinear relationship.

Regression analysis describes the relationship between two variables. A requirement is that this relationship be linear. The same is true of the correlation coefficient. It measures the strength of a linear relationship between two variables. But what if the relationship is not linear? The remedy is to rescale one or both of the variables so the new relationship is linear. For example, instead of using the actual values of the dependent variable, y, we would create a new dependent variable by computing the log to the

base 10 of y, Log(y). This calculation is called a transformation. Other common transformations include taking the square root, taking the reciprocal, or squaring one or both of the variables.

Thus, two variables could be closely related, but their relationship is not linear. Be cautious when you are interpreting the correlation coefficient or a regression equation. These statistics may indicate there is no linear relationship, but there could be a relationship of some other nonlinear or curvilinear form. The following example explains the details.

EXAMPLE

GroceryLand Supermarkets is a regional grocery chain with over 300 stores located in the midwestern United States. The corporate director of marketing for GroceryLand wishes to study the effect of price on the weekly sales of two-liter bottles of their private-brand diet cola. The objectives of the study are:

1. To determine whether there is a relationship between selling price and weekly sales. Is this relationship direct or indirect? Is it strong or weak?
2. To determine the effect of price increases or decreases on sales. Can we effectively forecast sales based on the price?

SOLUTION

To begin the project, the marketing director meets with the vice president of sales and other company staff members. They decide that it would be reasonable to price the two-liter bottle of their private-brand diet cola from $0.50 up to $2.00. To collect the data needed to analyze the relationship between price and sales, the marketing director selects a random sample of 20 stores and then randomly assigns a selling price for the two-liter bottle of diet cola between $0.50 and $2.00 to each selected store. The director contacts each of the 20 store managers included in the study to tell them the selling price and ask them to report the sales for the product at the end of the week. The results are reported below. For example, store number A-17 sold 181 two-liter bottles of diet cola at $0.50 each.

GroceryLand Sales and Price Data

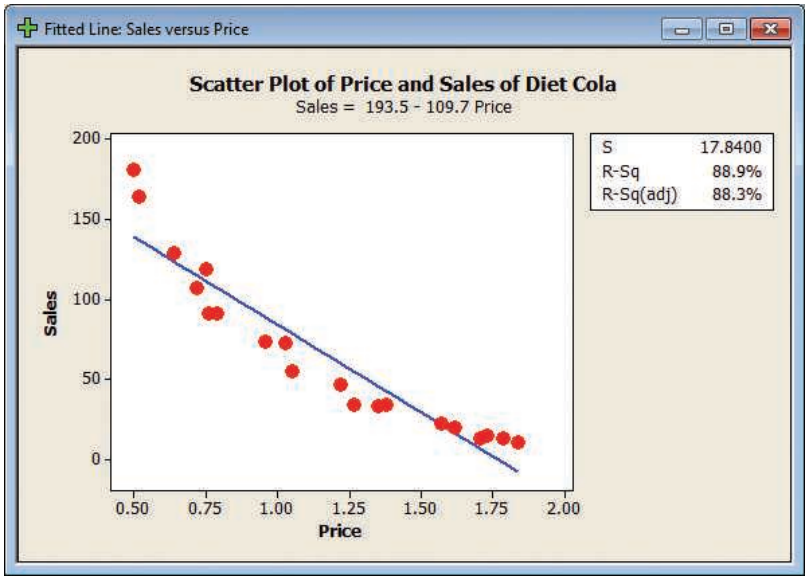
Store Number	Price	Sales
A-17	0.50	181
A-121	1.35	33
A-227	0.79	91
A-135	1.71	13
A-6	1.38	34
A-282	1.22	47
A-172	1.03	73
A-296	1.84	11
A-143	1.73	15
A-66	1.62	20

GroceryLand Sales and Price Data

Store Number	Price	Sales
A-30	0.76	91
A-127	1.79	13
A-266	1.57	22
A-117	1.27	34
A-132	0.96	74
A-120	0.52	164
A-272	0.64	129
A-120	1.05	55
A-194	0.72	107
A-105	0.75	119

To examine the relationship between Price and Sales, we use regression analysis setting *Price* as the independent variable and *Sales* as the dependent

variable. The analysis will provide important information about the relationship between the variables. The analysis is summarized in the following Minitab output.

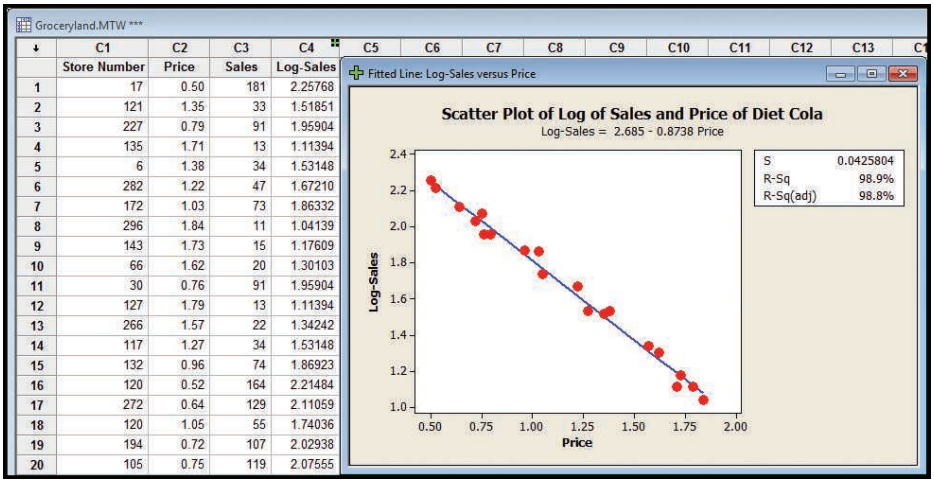


From the output, we can make these conclusions:

1. The relationship between the two variables is inverse or indirect. As the *Price* of the cola increases, the *Sales* of the product decreases. Given basic economic theory of price and demand, this is expected.
2. There is a strong relationship between the two variables. The coefficient of determination is 88.9%. So 88.9% of the variation in *Sales* is accounted for by the variation in *Price*. From the coefficient of determination, we can compute the correlation coefficient as the square root of the coefficient of determination. The correlation coefficient is the square root of 0.889, or 0.943. The sign of the correlation coefficient is negative because sales are inversely related to price. Therefore, the correlation coefficient is −0.943.
3. Before continuing our summary of conclusions, we should look carefully at the scatter diagram and the plot of the regression line. The assumption of a linear relationship is tenuous. If the relationship is linear, the data points should be distributed both above and below the line over the entire range of the independent variable. However, for the highest and lowest prices, the data points are above the regression line. For the selling prices in the middle, most of the data points are below the regression line. So the linear regression equation does not effectively describe the relationship between *Price* and *Sales*. A transformation of the data is needed to create a linear relationship.

By transforming one of the variables, we may be able to change the nonlinear relationship between the variables to a linear relationship. Of the possible choices, the director of marketing decides to transform the dependent variable, *Sales,* by taking the logarithm to the base 10 of each *Sales* value. Note the new variable, *Log-Sales,* in the following analysis. Now, the regression analysis uses *Log-Sales* as

the dependent variable and *Price* as the independent variable. This analysis is reported below.



	C1	C2	C3	C4
	Store Number	Price	Sales	Log-Sales
1	17	0.50	181	2.25768
2	121	1.35	33	1.51851
3	227	0.79	91	1.95904
4	135	1.71	13	1.11394
5	6	1.38	34	1.53148
6	282	1.22	47	1.67210
7	172	1.03	73	1.86332
8	296	1.84	11	1.04139
9	143	1.73	15	1.17609
10	66	1.62	20	1.30103
11	30	0.76	91	1.95904
12	127	1.79	13	1.11394
13	266	1.57	22	1.34242
14	117	1.27	34	1.53148
15	132	0.96	74	1.86923
16	120	0.52	164	2.21484
17	272	0.64	129	2.11059
18	120	1.05	55	1.74036
19	194	0.72	107	2.02938
20	105	0.75	119	2.07555

What can we conclude from the regression analysis using the transformation of the dependent variable *Sales*?

1. By transforming the dependent variable, *Sales,* we increase the coefficient of determination from 0.889 to 0.989. So *Price* now explains nearly all of the variation in *Log-Sales*.
2. Compare this result with the scatter diagram before we transformed the dependent variable. The transformed data seem to fit the linear relationship requirement much better. Observe that the data points are both above and below the regression line over the range of *Price*.
3. The regression equation is $\hat{y} = 2.685 - 0.8738x$. The sign of the slope value is negative, confirming the inverse association between the variables. We can use the new equation to estimate sales and study the effect of changes in price. For example, if we decided to sell the two-liter bottle of diet cola for $1.25, the predicted *Log-Sales* is:

$$\hat{y} = 2.685 - 0.8738x = 2.685 - 0.8738\,(1.25) = 1.593$$

Remember that the regression equation now predicts the log, base 10, of *Sales.* Therefore, we must undo the transformation by taking the antilog of 1.593, which is $10^{1.593}$, or 39.174. So, if we price the two-liter diet cola product at $1.25, the predicted weekly sales are 39 bottles. If we increase the price to $2.00, the regression equation would predict a value of .9374. Taking the antilog, $10^{.9374}$, the predicted sales decrease to 8.658, or, rounding, 9 two-liter bottles per week. Clearly, as price increases, sales decrease. This relationship will be very helpful to GroceryLand when making pricing decisions for this product.

EXERCISES

35. FILE Given the following sample of five observations, develop a scatter diagram, using x as the independent variable and y as the dependent variable, and compute the correlation coefficient. Does the relationship between the variables appear to be linear? Try squaring the x variable and then develop a scatter diagram and determine the correlation coefficient. Summarize your analysis.

x	−8	−16	12	2	18
y	58	247	153	3	341

36. FILE Every April, The Masters—one of the most prestigious golf tournaments on the PGA golf tour—is played in Augusta, Georgia. In 2016, 55 players received prize money. The 2016 winner, Danny Willett, earned a prize of $1,800,000. Jordan Spieth and Lee Westwood tied for second place, earning $880,000. Two amateur players finished "in the money" but they could not accept the prize money. They are not included in the data. The data are briefly summarized below. Each player has three corresponding variables: finishing position, score, and prize (in dollars). We want to study the relationship between score and prize.

Position	Player	Score	Prize
1	Danny Willett	283	$1,800,000
2(tied)	Jordan Spieth	286	$880,000
2(tied)	Lee Westwood	286	$880,000
4(tied)	Paul Casey	287	$413,333
4(tied)	J.B. Holmes	287	$413,333
4(tied)	Dustin Johnson	287	$413,333
..	..	..	..
..	..	..	..
..	..	..	..
52(tied)	Keegan Bradley	301	$24,900
52(tied)	Larry Mize	301	$24,900
54	Hunter Mahan	302	$24,000
55(tied)	Kevin Na	303	$23,400
55(tied)	Cameron Smith	303	$23,400
57	Thongchai Jaidee	307	$23,000

a. Using *Score* as the independent variable and *Prize* as the dependent variable, develop a scatter diagram. Does the relationship appear to be linear? Does it seem reasonable that as *Score* increases the *Prize* decreases?
b. What percentage of the variation in the dependent variable, *Prize,* is accounted for by the independent variable, *Score*?
c. Calculate a new variable, *Log-Prize*, computing the log to the base 10 of *Prize*. Draw a scatter diagram with *Log-Prize* as the dependent variable and *Score* as the independent variable.
d. Develop a regression equation and compute the coefficient of determination using *Log-Prize* as the dependent variable.
e. Compare the coefficient of determination in parts (b) and (d). What do you conclude?
f. Write out the regression equation developed in part (d). If a player shot an even par score of 288 for the four rounds, how much would you expect that player to earn?

CHAPTER SUMMARY

I. A scatter diagram is a graphic tool used to portray the relationship between two variables.
  A. The dependent variable is scaled on the *Y*-axis and is the variable being estimated.
  B. The independent variable is scaled on the *X*-axis and is the variable used as the predictor.
II. The correlation coefficient measures the strength of the linear association between two variables.
  A. Both variables must be at least the interval scale of measurement.
  B. The correlation coefficient can range from −1.00 to 1.00.

C. If the correlation between the two variables is 0, there is no association between them.
D. A value of 1.00 indicates perfect positive correlation, and a value of −1.00 indicates perfect negative correlation.
E. A positive sign means there is a direct relationship between the variables, and a negative sign means there is an inverse relationship.
F. It is designated by the letter r and found by the following equation:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} \qquad \textbf{[13–1]}$$

G. To test a hypothesis that a population correlation is different from 0, we use the following statistic:

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}} \qquad \text{with } n - 2 \text{ degrees of freedom} \qquad \textbf{[13–2]}$$

III. In regression analysis, we estimate one variable based on another variable.
A. The variable being estimated is the dependent variable.
B. The variable used to make the estimate or predict the value is the independent variable.
1. The relationship between the variables is linear.
2. Both the independent and the dependent variables must be interval or ratio scale.
3. The least squares criterion is used to determine the regression equation.
IV. The least squares regression line is of the form $\hat{y} = a + bx$
A. $\hat{y}$ is the estimated value of y for a selected value of x.
B. a is the constant or intercept.
1. It is the value of $\hat{y}$ when $x = 0$.
2. a is computed using the following equation.

$$a = \bar{y} - b\bar{x} \qquad \textbf{[13–5]}$$

C. b is the slope of the fitted line.
1. It shows the amount of change in $\hat{y}$ for a change of one unit in x.
2. A positive value for b indicates a direct relationship between the two variables. A negative value indicates an inverse relationship.
3. The sign of b and the sign of r, the correlation coefficient, are always the same.
4. b is computed using the following equation.

$$b = r\left(\frac{s_y}{s_x}\right) \qquad \textbf{[13–4]}$$

D. x is the value of the independent variable.
V. For a regression equation, the slope is tested for significance.
A. We test the hypothesis that the slope of the line in the population is 0.
1. If we do not reject the null hypothesis, we conclude there is no relationship between the two variables.
2. The test is equivalent to the test for the correlation coefficient.
B. When testing the null hypothesis about the slope, the test statistic is with $n - 2$ degrees of freedom:

$$t = \frac{b - 0}{s_b} \qquad \textbf{[13–6]}$$

VI. The standard error of estimate measures the variation around the regression line.
A. It is in the same units as the dependent variable.
B. It is based on squared deviations from the regression line.
C. Small values indicate that the points cluster closely about the regression line.
D. It is computed using the following formula.

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} \qquad \textbf{[13–7]}$$

VII. The coefficient of determination is the proportion of the variation of a dependent variable explained by the independent variable.
- **A.** It ranges from 0 to 1.0.
- **B.** It is the square of the correlation coefficient.
- **C.** It is found from the following formula.

$$r^2 = \frac{\text{SSR}}{\text{SS Total}} = 1 - \frac{\text{SSE}}{\text{SS Total}} \qquad \textbf{[13–8]}$$

VIII. Inference about linear regression is based on the following assumptions.
- **A.** For a given value of x, the values of Y are normally distributed about the line of regression.
- **B.** The standard deviation of each of the normal distributions is the same for all values of x and is estimated by the standard error of estimate.
- **C.** The deviations from the regression line are independent, with no pattern to the size or direction.

IX. There are two types of interval estimates.
- **A.** In a confidence interval, the mean value of y is estimated for a given value of x.
  - **1.** It is computed from the following formula.

$$\hat{y} \pm ts_{y \cdot x}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}} \qquad \textbf{[13–11]}$$

  - **2.** The width of the interval is affected by the level of confidence, the size of the standard error of estimate, and the size of the sample, as well as the value of the independent variable.
- **B.** In a prediction interval, the individual value of y is estimated for a given value of x.
  - **1.** It is computed from the following formula.

$$\hat{y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}} \qquad \textbf{[13–12]}$$

  - **2.** The difference between formulas (13–11) and (13–12) is the 1 under the radical.
    - **a.** The prediction interval will be wider than the confidence interval.
    - **b.** The prediction interval is also based on the level of confidence, the size of the standard error of estimate, the size of the sample, and the value of the independent variable.

PRONUNCIATION KEY

SYMBOL	MEANING	PRONUNCIATION
Σxy	Sum of the products of x and y	*Sum x y*
ρ	Correlation coefficient in the population	*Rho*
$\hat{y}$	Estimated value of Y	*y hat*
$s_{y \cdot x}$	Standard error of estimate	*s sub y dot x*
r^2	Coefficient of determination	*r square*

CHAPTER EXERCISES

37. A regional commuter airline selected a random sample of 25 flights and found that the correlation between the number of passengers and the total weight, in pounds, of luggage stored in the luggage compartment is 0.94. Using the .05 significance level, can we conclude that there is a positive association between the two variables?

38. A sociologist claims that the success of students in college (measured by their GPA) is related to their family's income. For a sample of 20 students, the correlation coefficient is 0.40. Using the 0.01 significance level, can we conclude that there is a positive correlation between the variables?

39. An Environmental Protection Agency study of 12 automobiles revealed a correlation of 0.47 between engine size and emissions. At the .01 significance level, can we conclude that there is a positive association between these variables? What is the *p*-value? Interpret.

40. FILE A suburban hotel derives its revenue from its hotel and restaurant operations. The owners are interested in the relationship between the number of rooms occupied on a nightly basis and the revenue per day in the restaurant. Below is a sample of 25 days (Monday through Thursday) from last year showing the restaurant income and number of rooms occupied.

Day	Revenue	Occupied	Day	Revenue	Occupied
1	$1,452	23	14	$1,425	27
2	1,361	47	15	1,445	34
3	1,426	21	16	1,439	15
4	1,470	39	17	1,348	19
5	1,456	37	18	1,450	38
6	1,430	29	19	1,431	44
7	1,354	23	20	1,446	47
8	1,442	44	21	1,485	43
9	1,394	45	22	1,405	38
10	1,459	16	23	1,461	51
11	1,399	30	24	1,490	61
12	1,458	42	25	1,426	39
13	1,537	54			

Use a statistical software package to answer the following questions.

a. Does the revenue seem to increase as the number of occupied rooms increases? Draw a scatter diagram to support your conclusion.

b. Determine the correlation coefficient between the two variables. Interpret the value.

c. Is it reasonable to conclude that there is a positive relationship between revenue and occupied rooms? Use the .10 significance level.

d. What percent of the variation in revenue in the restaurant is accounted for by the number of rooms occupied?

41. FILE The table below shows the number of cars (in millions) sold in the United States for various years and the percent of those cars manufactured by GM.

Year	Cars Sold (millions)	Percent GM	Year	Cars Sold (millions)	Percent GM
1950	6.0	50.2	1985	15.4	40.1
1955	7.8	50.4	1990	13.5	36.0
1960	7.3	44.0	1995	15.5	31.7
1965	10.3	49.9	2000	17.4	28.6
1970	10.1	39.5	2005	16.9	26.9
1975	10.8	43.1	2010	11.6	19.1
1980	11.5	44.0	2015	17.5	17.6

Use a statistical software package to answer the following questions.

a. Is the number of cars sold directly or indirectly related to GM's percentage of the market? Draw a scatter diagram to show your conclusion.

b. Determine the correlation coefficient between the two variables. Interpret the value.

c. Is it reasonable to conclude that there is a negative association between the two variables? Use the .01 significance level.

d. How much of the variation in GM's market share is accounted for by the variation in cars sold?

42. For a sample of 40 large U.S. cities, the correlation between the mean number of square feet per office worker and the mean monthly rental rate in the central business district is −0.363. At the .05 significance level, can we conclude that there is a negative association between the two variables.

43. FILE For each of the 32 National Football League teams, the numbers of points scored and allowed during the 2016 season are shown below.

TEAM	Conference	PTS Scored	PTS Allowed	TEAM	Conference	PTS Scored	PTS Allowed
Baltimore	AFC	343	321	Arizona	NFC	418	362
Buffalo	AFC	399	378	Atlanta	NFC	540	406
Cincinnati	AFC	325	315	Carolina	NFC	369	402
Cleveland	AFC	264	452	Chicago	NFC	279	399
Denver	AFC	333	297	Dallas	NFC	421	306
Houston	AFC	279	328	Detroit	NFC	346	358
Indianapolis	AFC	411	392	Green Bay	NFC	432	388
Jacksonville	AFC	318	400	Los Angeles	NFC	224	394
Kansas City	AFC	389	311	Minnesota	NFC	327	307
Miami	AFC	363	380	NY Giants	NFC	469	454
New England	AFC	441	250	New Orleans	NFC	310	284
NY Jets	AFC	275	409	Philadelphia	NFC	367	331
Oakland	AFC	416	385	San Francisco	NFC	309	480
Pittsburgh	AFC	399	327	Seattle	NFC	354	292
San Diego	AFC	410	423	Tampa Bay	NFC	354	369
Tennessee	AFC	381	378	Washington	NFC	396	383

Assuming these are sample data, answer the following questions. You may use statistical software to assist you.

a. What is the correlation coefficient between these variables? Are you surprised the association is negative? Interpret your results.

b. Find the coefficient of determination. What does it say about the relationship?

c. At the .05 significance level, can you conclude there is a negative association between "points scored" and "points allowed"?

d. At the .05 significance level, can you conclude there is a negative association between "points scored" and "points allowed" for each conference?

44. FILE The Cotton Mill is an upscale chain of women's clothing stores, located primarily in the southwest United States. Due to recent success, The Cotton Mill's top management is planning to expand by locating new stores in other regions of the country. The director of planning has been asked to study the relationship between yearly sales and the store size. As part of the study, the director selects a sample of 25 stores and determines the size of the store in square feet and the sales for last year. The sample data follow. The use of statistical software is suggested.

Store Size (thousands of square feet)	Sales (millions $)	Store Size (thousands of square feet)	Sales (millions $)
3.7	9.18	0.4	0.55
2.0	4.58	4.2	7.56
5.0	8.22	3.1	2.23
0.7	1.45	2.6	4.49
2.6	6.51	5.2	9.90
2.9	2.82	3.3	8.93
5.2	10.45	3.2	7.60
5.9	9.94	4.9	3.71
3.0	4.43	5.5	5.47
2.4	4.75	2.9	8.22
2.4	7.30	2.2	7.17
0.5	3.33	2.3	4.35
5.0	6.76		

a. Draw a scatter diagram. Use store size as the independent variable. Does there appear to be a relationship between the two variables. Is it positive or negative?
b. Determine the correlation coefficient and the coefficient of determination. Is the relationship strong or weak? Why?
c. At the .05 significance level, can we conclude there is a significant positive correlation?

45. FILE The manufacturer of Cardio Glide exercise equipment wants to study the relationship between the number of months since the glide was purchased and the time, in hours, the equipment was used last week.

Person	Months Owned	Hours Exercised	Person	Months Owned	Hours Exercised
Rupple	12	4	Massa	2	8
Hall	2	10	Sass	8	3
Bennett	6	8	Karl	4	8
Longnecker	9	5	Malrooney	10	2
Phillips	7	5	Veights	5	5

a. Plot the information on a scatter diagram. Let hours of exercise be the dependent variable. Comment on the graph.
b. Determine the correlation coefficient. Interpret.
c. At the .01 significance level, can we conclude that there is a negative association between the variables?

46. The following regression equation was computed from a sample of 20 observations:

$$\hat{y} = 15 - 5x$$

SSE was found to be 100 and SS total, 400.
a. Determine the standard error of estimate.
b. Determine the coefficient of determination.
c. Determine the correlation coefficient. (Caution: Watch the sign!)

47. FILE City planners believe that larger cities are populated by older residents. To investigate the relationship, data on population and median age in 10 large cities were collected.

City	Population City (in millions)	Median Age
Chicago, IL	2.833	31.5
Dallas, TX	1.233	30.5
Houston, TX	2.144	30.9
Los Angeles, CA	3.849	31.6
New York, NY	8.214	34.2
Philadelphia, PA	1.448	34.2
Phoenix, AZ	1.513	30.7
San Antonio, TX	1.297	31.7
San Diego, CA	1.257	32.5
San Jose, CA	0.930	32.6

a. Plot these data on a scatter diagram with median age as the dependent variable.
b. Find the correlation coefficient.
c. A regression analysis was performed and the resulting regression equation is Median Age = 31.4 + 0.272 Population. Interpret the meaning of the slope.
d. Estimate the median age for a city of 2.5 million people.
e. Here is a portion of the regression software output. What does it tell you?

```
Predictor          Coef     SE Coef         T         P
Constant        31.3672      0.6158     50.94     0.000
Population       0.2722      0.1901      1.43     0.190
```

f. Using the .10 significance level, test the significance of the slope. Interpret the result. Is there a significant relationship between the two variables?

48. FILE Emily Smith decides to buy a fuel-efficient used car. Here are several vehicles she is considering, with the estimated cost to purchase and the age of the vehicle.

Vehicle	Estimated Cost	Age
Honda Insight	$5,555	8
Toyota Prius	$17,888	3
Toyota Prius	$9,963	6
Toyota Echo	$6,793	5
Honda Civic Hybrid	$10,774	5
Honda Civic Hybrid	$16,310	2
Chevrolet Cruz	$2,475	8
Mazda3	$2,808	10
Toyota Corolla	$7,073	9
Acura Integra	$8,978	8
Scion xB	$11,213	2
Scion xA	$9,463	3
Mazda3	$15,055	2
Mini Cooper	$20,705	2

a. Plot these data on a scatter diagram with estimated cost as the dependent variable.
b. Find the correlation coefficient.
c. A regression analysis was performed and the resulting regression equation is Estimated Cost = 18358 − 1534 Age. Interpret the meaning of the slope.
d. Estimate the cost of a five-year-old car.
e. Here is a portion of the regression software output. What does it tell you?

Predictor	Coef	SE Coef	T	P
Constant	18358	1817	10.10	0.000
Age	−1533.6	306.3	−5.01	0.000

f. Using the .10 significance level, test the significance of the slope. Interpret the result. Is there a significant relationship between the two variables?

49. FILE The National Highway Association is studying the relationship between the number of bidders on a highway project and the winning (lowest) bid for the project. Of particular interest is whether the number of bidders increases or decreases the amount of the winning bid.

Project	Number of Bidders, *x*	Winning Bid ($ millions), *y*	Project	Number of Bidders, *x*	Winning Bid ($ millions), *y*
1	9	5.1	9	6	10.3
2	9	8.0	10	6	8.0
3	3	9.7	11	4	8.8
4	10	7.8	12	7	9.4
5	5	7.7	13	7	8.6
6	10	5.5	14	7	8.1
7	7	8.3	15	6	7.8
8	11	5.5			

a. Determine the regression equation. Interpret the equation. Do more bidders tend to increase or decrease the amount of the winning bid?
b. Estimate the amount of the winning bid if there were seven bidders.
c. A new entrance is to be constructed on the Ohio Turnpike. There are seven bidders on the project. Develop a 95% prediction interval for the winning bid.
d. Determine the coefficient of determination. Interpret its value.

50. FILE Mr. William Profit is studying companies going public for the first time. He is particularly interested in the relationship between the size of the offering and the price per share. A sample of 15 companies that recently went public revealed the following information.

Company	Size ($ millions), *x*	Price per Share, *y*	Company	Size ($ millions), *x*	Price per Share, *y*
1	9.0	10.8	9	160.7	11.3
2	94.4	11.3	10	96.5	10.6
3	27.3	11.2	11	83.0	10.5
4	179.2	11.1	12	23.5	10.3
5	71.9	11.1	13	58.7	10.7
6	97.9	11.2	14	93.8	11.0
7	93.5	11.0	15	34.4	10.8
8	70.0	10.7			

a. Determine the regression equation.
b. Conduct a test to determine whether the slope of the regression line is positive.
c. Determine the coefficient of determination. Do you think Mr. Profit should be satisfied with using the size of the offering as the independent variable?

51. FILE Bardi Trucking Co., located in Cleveland, Ohio, makes deliveries in the Great Lakes region, the Southeast, and the Northeast. Jim Bardi, the president, is studying the relationship between the distance a shipment must travel and the length of time, in days, it takes the shipment to arrive at its destination. To investigate, Mr. Bardi selected a random sample of 20 shipments made last month. Shipping distance is the independent variable and shipping time is the dependent variable. The results are as follows:

Shipment	Distance (miles)	Shipping Time (days)	Shipment	Distance (miles)	Shipping Time (days)
1	656	5	11	862	7
2	853	14	12	679	5
3	646	6	13	835	13
4	783	11	14	607	3
5	610	8	15	665	8
6	841	10	16	647	7
7	785	9	17	685	10
8	639	9	18	720	8
9	762	10	19	652	6
10	762	9	20	828	10

a. Draw a scatter diagram. Based on these data, does it appear that there is a relationship between how many miles a shipment has to go and the time it takes to arrive at its destination?
b. Determine the correlation coefficient. Can we conclude that there is a positive correlation between distance and time? Use the .05 significance level.
c. Determine and interpret the coefficient of determination.
d. Determine the standard error of estimate.
e. Would you recommend using the regression equation to predict shipping time? Why or why not.

52. FILE Super Markets Inc. is considering expanding into the Scottsdale, Arizona, area. You, as director of planning, must present an analysis of the proposed expansion to the operating committee of the board of directors. As a part of your proposal, you need to include information on the amount people in the region spend per month for grocery items. You would also like to include information on the relationship between the amount spent for grocery items and income. Your assistant gathered the following sample information.

Household	Amount Spent	Monthly Income
1	$ 555	$4,388
2	489	4,558
⋮	⋮	⋮
39	1,206	9,862
40	1,145	9,883

a. Let the amount spent be the dependent variable and monthly income the independent variable. Create a scatter diagram, using a software package.
b. Determine the regression equation. Interpret the slope value.
c. Determine the correlation coefficient. Can you conclude that it is greater than 0?

53. FILE Below is information on the price per share and the dividend for a sample of 30 companies.

Company	Price per Share	Dividend
1	$20.00	$ 3.14
2	22.01	3.36
⋮	⋮	⋮
29	77.91	17.65
30	80.00	17.36

a. Calculate the regression equation using selling price based on the annual dividend.
b. Test the significance of the slope.
c. Determine the coefficient of determination. Interpret its value.
d. Determine the correlation coefficient. Can you conclude that it is greater than 0 using the .05 significance level?

54. A highway employee performed a regression analysis of the relationship between the number of construction work-zone fatalities and the number of unemployed people in a state. The regression equation is Fatalities = 12.7 + 0.000114 (Unemp). Some additional output is:

```
Predictor          Coef       SE Coef       T       P
Constant         12.726         8.115    1.57   0.134
Unemp        0.00011386    0.00002896    3.93   0.001
Analysis of Variance
Source             DF       SS       MS       F       P
Regression          1    10354    10354   15.46   0.001
Residual Error     18    12054      670
Total              19    22408
```

a. How many states were in the sample?
b. Determine the standard error of estimate.
c. Determine the coefficient of determination.
d. Determine the correlation coefficient.
e. At the .05 significance level, does the evidence suggest there is a positive association between fatalities and the number unemployed?

55. A regression analysis relating the current market value in dollars to the size in square feet of homes in Greene County, Tennessee, follows. The regression equation is: Value = −37,186 + 65.0 Size.

```
Predictor          Coef       SE Coef       T       P
Constant         −37186          4629   −8.03   0.000
Size            64.993         3.047   21.33   0.000
Analysis of Variance
Source             DF            SS            MS        F       P
Regression          1   13548662082   13548662082   454.98   0.000
Residual Error     33     982687392      29778406
Total              34   14531349474
```

a. How many homes were in the sample?
b. Compute the standard error of estimate.
c. Compute the coefficient of determination.
d. Compute the correlation coefficient.
e. At the .05 significance level, does the evidence suggest a positive association between the market value of homes and the size of the home in square feet?

56. FILE The following table shows the mean annual percent return on capital (profitability) and the mean annual percentage sales growth for eight aerospace and defense companies.

Company	Profitability	Growth
Alliant Techsystems	23.1	8.0
Boeing	13.2	15.6
General Dynamics	24.2	31.2
Honeywell	11.1	2.5
L-3 Communications	10.1	35.4
Northrop Grumman	10.8	6.0
Rockwell Collins	27.3	8.7
United Technologies	20.1	3.2

a. Compute the correlation coefficient. Conduct a test of hypothesis to determine if it is reasonable to conclude that the population correlation is greater than zero. Use the .05 significance level.
b. Develop the regression equation for profitability based on growth. Can we conclude that the slope of the regression line is negative?
c. Use a software package to determine the residual for each observation. Which company has the largest residual?

57. FILE The following data show the retail price for 12 randomly selected laptop computers along with their corresponding processor speeds in gigahertz.

Computer	Speed	Price	Computer	Speed	Price
1	2.0	1008.50	7	2.0	1098.50
2	1.6	461.00	8	1.6	693.50
3	1.6	532.00	9	2.0	1057.00
4	1.8	971.00	10	1.6	1001.00
5	2.0	1068.50	11	1.0	468.50
6	1.2	506.00	12	1.4	434.50

a. Develop a linear equation that can be used to describe how the price depends on the processor speed.
b. Based on your regression equation, is there one machine that seems particularly over- or underpriced?
c. Compute the correlation coefficient between the two variables. At the .05 significance level, conduct a test of hypothesis to determine if the population correlation is greater than zero.

58. FILE A consumer buying cooperative tested the effective heating area of 20 different electric space heaters with different wattages. Here are the results.

Heater	Wattage	Area	Heater	Wattage	Area
1	1,500	205	11	1,250	116
2	750	70	12	500	72
3	1,500	199	13	500	82
4	1,250	151	14	1,500	206
5	1,250	181	15	2,000	245
6	1,250	217	16	1,500	219
7	1,000	94	17	750	63
8	2,000	298	18	1,500	200
9	1,000	135	19	1,250	151
10	1,500	211	20	500	44

a. Compute the correlation between the wattage and heating area. Is there a direct or an indirect relationship?
b. Conduct a test of hypothesis to determine if it is reasonable that the coefficient is greater than zero. Use the .05 significance level.
c. Develop the regression equation for effective heating based on wattage.
d. Which heater looks like the "best buy" based on the size of the residual?

59. FILE A dog trainer is exploring the relationship between the size of the dog (weight in pounds) and its daily food consumption (measured in standard cups). Below is the result of a sample of 18 observations.

Dog	Weight	Consumption	Dog	Weight	Consumption
1	41	3	10	91	5
2	148	8	11	109	6
3	79	5	12	207	10
4	41	4	13	49	3
5	85	5	14	113	6
6	111	6	15	84	5
7	37	3	16	95	5
8	111	6	17	57	4
9	41	3	18	168	9

a. Compute the correlation coefficient. Is it reasonable to conclude that the correlation in the population is greater than zero? Use the .05 significance level.
b. Develop the regression equation for cups based on the dog's weight. How much does each additional cup change the estimated weight of the dog?
c. Is one of the dogs a big undereater or overeater?

60. Waterbury Insurance Company wants to study the relationship between the amount of fire damage and the distance between the burning house and the nearest fire station. This information will be used in setting rates for insurance coverage. For a sample of 30 claims for the last year, the director of the actuarial department determined the distance from the fire station (x) and the amount of fire damage, in thousands of dollars (y). The MegaStat output is reported below.

```
ANOVA table
Source                SS    df           MS        F
Regression     1,864.5782     1   1,864.5782    38.83
Residual       1,344.4934    28      48.0176
Total          3,209.0716    29
Regression output
Variables     Coefficients    Std. Error  t(df = 28)
Intercept          12.3601        3.2915       3.755
Distance–X          4.7956        0.7696       6.231
```

Answer the following questions.
a. Write out the regression equation. Is there a direct or indirect relationship between the distance from the fire station and the amount of fire damage?
b. How much damage would you estimate for a fire 5 miles from the nearest fire station?
c. Determine and interpret the coefficient of determination.
d. Determine the correlation coefficient. Interpret its value. How did you determine the sign of the correlation coefficient?
e. Conduct a test of hypothesis to determine if there is a significant relationship between the distance from the fire station and the amount of damage. Use the .01 significance level and a two-tailed test.

61. FILE TravelAir.com samples domestic airline flights to explore the relationship between airfare and distance. The service would like to know if there is a correlation between airfare and flight distance. If there is a correlation, what percentage of the variation in airfare is accounted for by distance? How much does each additional mile add to the fare? The data follow.

Origin	Destination	Distance	Fare
Detroit, MI	Myrtle Beach, SC	636	$109
Baltimore, MD	Sacramento, CA	2,395	252
Las Vegas, NV	Philadelphia, PA	2,176	221
Sacramento, CA	Seattle, WA	605	151
Atlanta, GA	Orlando, FL	403	138
Boston, MA	Miami, FL	1,258	209
Chicago, IL	Covington, KY	264	254
Columbus, OH	Minneapolis, MN	627	259
Fort Lauderdale, FL	Los Angeles, CA	2,342	215
Chicago, IL	Indianapolis, IN	177	128
Philadelphia, PA	San Francisco, CA	2,521	348
Houston, TX	Raleigh/Durham, NC	1,050	224
Houston, TX	Midland/Odessa, TX	441	175
Cleveland, OH	Dallas/Ft.Worth, TX	1,021	256
Baltimore, MD	Columbus, OH	336	121
Boston, MA	Covington, KY	752	252
Kansas City, MO	San Diego, CA	1,333	206
Milwaukee, WI	Phoenix, AZ	1,460	167
Portland, OR	Washington, DC	2,350	308
Phoenix, AZ	San Jose, CA	621	152
Baltimore, MD	St. Louis, MO	737	175
Houston, TX	Orlando, FL	853	191
Houston, TX	Seattle, WA	1,894	231
Burbank, CA	New York, NY	2,465	251
Atlanta, GA	San Diego, CA	1,891	291
Minneapolis, MN	New York, NY	1,028	260
Atlanta, GA	West Palm Beach, FL	545	123
Kansas City, MO	Seattle, WA	1,489	211
Baltimore, MD	Portland, ME	452	139
New Orleans, LA	Washington, DC	969	243

a. Draw a scatter diagram with *Distance* as the independent variable and *Fare* as the dependent variable. Is the relationship direct or indirect?
b. Compute the correlation coefficient. At the .05 significance level, is it reasonable to conclude that the correlation coefficient is greater than zero?
c. What percentage of the variation in *Fare* is accounted for by *Distance* of a flight?
d. Determine the regression equation. How much does each additional mile add to the fare? Estimate the fare for a 1,500-mile flight.
e. A traveler is planning to fly from Atlanta to London Heathrow. The distance is 4,218 miles. She wants to use the regression equation to estimate the fare. Explain why it would not be a good idea to estimate the fare for this international flight with the regression equation.

DATA ANALYTICS

62. FILE The North Valley Real Estate data reports information on homes on the market.
a. Let selling price be the dependent variable and size of the home the independent variable. Determine the regression equation. Estimate the selling price for a home with an area of 2,200 square feet. Determine the 95% confidence interval for all 2,200 square foot homes and the 95% prediction interval for the selling price of a home with 2,200 square feet.

b. Let days-on-the-market be the dependent variable and price be the independent variable. Determine the regression equation. Estimate the days-on-the-market of a home that is priced at $300,000. Determine the 95% confidence interval of days-on-the-market for homes with a mean price of $300,000, and the 95% prediction interval of days-on-the-market for a home priced at $300,000.

c. Can you conclude that the independent variables "days on the market" and "selling price" are positively correlated? Are the size of the home and the selling price positively correlated? Use the .05 significance level. Report the *p*-value of the test. Summarize your results in a brief report.

63. FILE Refer to the Baseball 2016 data, which reports information on the 2016 Major League Baseball season. Let attendance be the dependent variable and total team salary be the independent variable. Determine the regression equation and answer the following questions.

a. Draw a scatter diagram. From the diagram, does there seem to be a direct relationship between the two variables?

b. What is the expected attendance for a team with a salary of $100.0 million?

c. If the owners pay an additional $30 million, how many more people could they expect to attend?

d. At the .05 significance level, can we conclude that the slope of the regression line is positive? Conduct the appropriate test of hypothesis.

e. What percentage of the variation in attendance is accounted for by salary?

f. Determine the correlation between attendance and team batting average and between attendance and team ERA. Which is stronger? Conduct an appropriate test of hypothesis for each set of variables.

64. FILE Refer to the LincolnVille School bus data. Develop a regression equation that expresses the relationship between age of the bus and maintenance cost. The age of the bus is the independent variable.

a. Draw a scatter diagram. What does this diagram suggest as to the relationship between the two variables? Is it direct or indirect? Does it appear to be strong or weak?

b. Develop a regression equation. How much does an additional year add to the maintenance cost. What is the estimated maintenance cost for a 10-year-old bus?

c. Conduct a test of hypothesis to determine whether the slope of the regression line is greater than zero. Use the .05 significance level. Interpret your findings from parts (a), (b), and (c) in a brief report.